Evidence-b(i)ased psychiatry

de Vries, Ymkje Anna

Evidence-b(i)ased psychiatry

Ymkje Anna de Vries

Printed by Ridderprint B.V.

Publication of this dissertation was financially supported by the University Medical Center Groningen and the University of Groningen.

ISBN: 978-94-034-0309-0 (printed version)
ISBN: 978-94-034-0308-3 (digital version)

university of groningen

Evidence-b(i)ased psychiatry

PhD thesis

to obtain the degree of PhD at the
University of Groningen
on the authority of the
Rector Magnificus Prof. E. Sterken
and in accordance with
the decision by the College of Deans.

This thesis will be defended in public on

Wednesday 21 February 2018 at 14.30 hours

by

Ymkje Anna de Vries

born on 18 July 1988
in Kollumerland en Nieuwkruisland

Supervisor
Prof. P. de Jonge

Co-supervisor
Dr. A.M. Roest

Assessment Committee
Prof. M.J. Postma
Prof. P.F.M. Verhaak
Prof. J. Spijker

Table of contents

Chapter 1

Introduction

Depression and anxiety

Major depressive disorder (MDD) and anxiety disorders are highly prevalent mental disorders. The lifetime risk for MDD has been estimated at 23.2%, while that of any anxiety disorder has been estimated at 31.5% [1]. The core symptoms of MDD are depressed mood and anhedonia (loss of pleasure or diminished interest) [2]. The anxiety disorders form a heterogeneous group, which consists of generalized anxiety disorder (GAD), social anxiety disorder (SAD), obsessive-compulsive disorder (OCD), post-traumatic stress disorder (PTSD), panic disorder, agoraphobia, and specific phobia [2]. Their shared feature is the presence of anxiety states, varying from worry to obsessions to panic attacks.

Although there are clear differences among anxiety disorders and between anxiety disorders and depression, for instance in age of onset [1] and episode duration [3], they are also highly comorbid [4], which may be due to an underlying 'internalizing' liability [5]. Anxiety disorders, due to their earlier age of onset, are often a precursor to later depression [6], but depression can also precede anxiety disorders [7, 8, 9] and is often accompanied by clinically significant anxiety even in those without an anxiety disorder [10].

Treatment approaches for these disorders are also similar. First-line treatment strategies for both MDD and anxiety disorders include antidepressants and psychotherapy, primarily cognitive-behavioral therapy (CBT) [11, 12, 13, 14, 15, 16]. Although there are distinct CBT approaches for different disorders (e.g. exposure and response prevention for the treatment of OCD [17]), trans-diagnostic approaches have also been developed [18]. While antidepressants were initially developed and approved for the treatment of depression, they were later found to have similar efficacy for anxiety disorders [19, 20, 21]. They are now preferred over benzodiazepines, which were previously used frequently for anxiety but which have fallen out of favor due to their potential for abuse and dependence, particularly with long-term use [22]. Currently, slightly more than half of all antidepressant prescriptions are for MDD, while about a quarter are for anxiety disorders [23]. Antidepressant use has greatly increased over the last two decades, with 12% of the American population [24] and nearly 6% of the Dutch population [25] now receiving antidepressants yearly.

This thesis

Hundreds, if not thousands, of randomized controlled trials (RCTs) have been performed over the last half-century to demonstrate the efficacy of antidepressants and psychotherapy for the treatment of MDD and anxiety disorders. In spite of this wealth of evidence, however, essential questions remain unanswered.

In the first place, the evidence base is threatened by the presence of reporting and citation biases. As a consequence of these biases, the true efficacy and safety of these treatments

remains unclear, and this thesis therefore aims to re-examine the evidence base in order to clarify the impact of bias on the (apparent) efficacy and safety of these treatments. For methodological reasons, the emphasis will be on antidepressants, although many of these biases apply equally regardless of treatment modality.

Secondly, these treatments appear to be only modestly effective. While some patients respond very well to the first treatment that is tried (whether antidepressants, psychotherapy, or a combination), others require multiple treatment trials before finally, through a process of trial and error, chancing upon a treatment that is effective for them, and still others do not seem to respond well to any treatment. Further complicating the picture, some patients also show a good response to placebo, suggesting that for a subset of people suffering from MDD or anxiety disorders, some combination of the passage of time, supportive conversations, and instilling hope for improvement may be sufficient, without the need for any active (and potentially harmful) treatment.

Therefore, the second aim of this thesis is to use the existing evidence base to investigate clinical predictors of treatment response, in order to identify those patients who are likely to benefit from receiving (active) treatment prior to initiating treatment or as early as possible in the course of treatment.

Evidence-based psychiatry: the problem of bias

Although antidepressant use has increased over the past decade, these medications have also been embroiled in controversy. Much of this controversy has been due to allegations that pharmaceutical companies buried the results of trials that did not show that the drug was effective or that suggested that the drug was associated with safety concerns [19, 26, 27, 28, 29].

One of these trials, known as Study 329, has become particularly infamous. This trial examined the efficacy and safety of paroxetine, a selective serotonin reuptake inhibitor (SSRI), compared to imipramine, an older tricyclic antidepressant, and placebo in depressed adolescents. Originally conducted in the 1990s and published in 2001 [30], the results of this trial were used by SmithKline Beecham (SKB, now GlaxoSmithKline), the manufacturer of paroxetine, to promote the use of paroxetine in young people, with claims that the results showed the "REMARKABLE efficacy and safety" of paroxetine in the treatment of adolescent depression [28].

However, from the beginning there were concerns that the published article misrepresented the results of the trial [31, 32]. This was done by confusing the pre-specified primary outcome (which did not show significant efficacy of paroxetine) with a post-hoc secondary outcome that did yield a statistically significant result for paroxetine. Furthermore, serious adverse events, including suicide attempts, that occurred at an increased rate in the paroxetine group were dismissed as being unrelated to the drug. Re-analysis of

the trial by independent investigators concluded that neither imipramine nor paroxetine showed efficacy for adolescent depression in this trial and that both drugs were associated with an increased rate of adverse events, including suicidal behavior for paroxetine and cardiovascular problems for imipramine [33].

No other psychiatric trial has achieved the same notoriety as Study 329. However, bias is actually ubiquitous and has been demonstrated within psychiatry, medical science as a whole, and science in general [34]. Biased reporting can be motivated by commercial interests, as in the case of Study 329, but it can also occur in the absence of any commercial interests. Although much attention has (deservedly) been devoted to bias in antidepressant trials, for instance, there is every reason to believe that bias is just as common in psychotherapy trials [35]. In general, the scientific community appears to be prejudiced against null findings, so-called negative results that do not, for example, show that an intervention is effective.

Initial concern about bias mainly revolved around the problem of study publication bias, which was first identified more than fifty years ago [36] and which occurs when the likelihood of publication of a study depends upon the direction or significance of the results [37, 38]. However, negative results face additional obstacles to getting as much visibility as positive results than just study publication bias. These other biases include (but are not limited to) outcome reporting bias, spin, and citation bias.

Outcome reporting bias occurs when pre-specified outcomes are omitted from the published article, when new outcomes are added without being identified as new, or when the status of (non-significant) primary and (significant) secondary outcomes are switched. Although outcome reporting bias probably occurs regardless of study design, it is most easily identified in RCTs, because other studies (e.g. observational studies) often do not have a protocol with clearly defined primary and secondary outcomes. Outcome reporting bias has been detected in 31 – 62% of published RCTs [38, 39], indicating the scope of the problem.

Spin is defined as specific reporting strategies, whether intentional or unintentional, that could distort the interpretation of results. For example, while an article with spin does report non-significant results on the primary outcome, it nevertheless concludes that the intervention is effective on the basis of statistically significant results on secondary outcomes, in subgroups, or in pre-post comparisons. Among a sample of RCTs with non-significant results on the primary outcome, 58% contained spin in the conclusions section of the abstract and 50% contained spin in the conclusions section of the article itself [40]. Spin can affect clinicians' interpretations of a trial, leading them to judge interventions as being more beneficial [41].

Finally, citation bias refers to the tendency to preferentially cite positive studies, which has been demonstrated in various literatures [42, 43, 44, 45, 46]. Since studies are more likely to be discovered when they are cited by other studies, citation bias serves to high-

light positive results while negative results may go unnoticed.

Within the literature on depression and anxiety disorders, a landmark study was published in 2008, showing that negative trials of antidepressants for MDD were much less likely to be published than positive trials [19]. Furthermore, when these negative trials were published, they were often published as if positive. In this study, 51% of all trials were positive, but in sharp contrast, 94% of *published* trials appeared to be positive, and the effect size of antidepressants was overestimated by 32%. Study publication bias is also present in psychotherapy trials. Within a cohort of National Institutes of Health-funded psychotherapy trials, 24% of all initiated trials remained unpublished, and the effect size of these unpublished trials was markedly lower than that of published trials, clearly showing a bias against publishing unfavorable findings [35].

While most research on bias has focused on the efficacy of treatments, some research has also examined safety. Adverse events are rarely monitored, let alone reported, in psychotherapy trials [47, 48, 49]. Consequently, we know very little about the possible adverse effects of psychotherapy, although it is clear that psychotherapy, like all effective treatments, can have negative as well as positive effects [50, 51, 52]. Adverse events in drug trials, however, are usually extensively monitored because of regulatory requirements, but reporting on these events is limited, both in psychiatry [53] and in other medical fields [54, 55, 56]. Previous research on two antidepressants (sertraline and duloxetine) has suggested that harm outcomes are poorly reported and serious adverse events in particular are not always reported fully or accurately [57, 58]. In general, however, relatively little attention has been devoted to harm outcomes compared to efficacy outcomes.

Precision psychiatry: who benefits from treatment?

The research on bias in antidepressant trials has shown that these medications are less effective and also less safe than previously thought, particularly in young people. However, although antidepressant efficacy may, on average, be more modest than expected or desired, it is likely that some people do experience a robust and clinically meaningful response to antidepressants, while others experience very little, or no benefit at all. In general, around 15% more people respond to antidepressants than to placebo [59]. Consequently, there is great interest in predicting who benefits from treatment, to distinguish between patients who will recover even without active treatment, patients who will benefit specifically from the treatment, and patients who may need a different or more intensive treatment to recover.

Much effort has been devoted to examining biological and genetic markers that could be associated with antidepressant response. There has long been a particular interest in the serotonin-transporter-linked polymorphic region (5-HTTLPR), for instance, because SSRIs, the most commonly used antidepressants, inhibit the reuptake of serotonin by

blocking the serotonin transporter. Meta-analyses on this topic, however, have come to diverging conclusions on whether 5-HTTLPR is actually associated with antidepressant response [60, 61].

In general, candidate gene studies like these (in which a specific gene of interest is studied) have proven unlikely to replicate reliably [62, 63], due to the unfortunate combination of very small sample sizes (increasing the likelihood that any statistically significant finding is a false positive), analytical flexibility (providing the possibility of selecting the analysis or outcome that worked "best", that is, yielded a statistically significant result [64]), and reporting bias. For that reason, the field of genetics has largely abandoned candidate gene studies in favor of genome-wide association studies (GWAS), which have proven to be much more reliable [65].

In the past decade, GWAS have demonstrated that genetic effects are usually extremely small and very large sample sizes are required to detect these effects [62]. Obtaining such sample sizes within the context of a treatment trial is a major challenge. A recent meta-analysis of antidepressant pharmacogenetics trials, which included 2,256 participants, found only one genome-wide significant association; a polygenic risk score accounted for only about 1% of the variance in treatment outcomes [66]. Hence, it seems unlikely that genetic research will soon deliver a major contribution to the prediction of antidepressant response.

Other biological markers may have more potential, but so far decades of research have not resulted in any markers that are sufficiently predictive to be useful [67]. Biological markers based on electroencephalograms (EEG), functional magnetic resonance imaging (fMRI), or the like are also unlikely to be feasible in routine clinical practice. Markers that can be derived from blood or saliva tests may be more feasible, but still require additional steps that are seldom performed in clinical practice. Consequently, the requirements of feasibility dictate that routinely obtained and readily available clinical information should be used whenever possible, in preference over much less efficient alternatives like neuroimaging, unless this is shown to be much more accurate [68].

One obvious clinical characteristic is disorder severity. A 2008 meta-analysis suggested that initial severity of depressive symptoms was associated with antidepressant response, such that people with milder depression experienced little benefit from taking antidepressants compared to placebo [69]. Other research appeared to confirm this [70, 71, 72], and clinical guidelines were updated to reflect this finding and to recommend against the use of antidepressants as a first-line treatment for mild depression [73, 74]. More recently, however, two large studies did not replicate this association [75, 76], casting doubt upon this finding.

Although antidepressants are also commonly used for anxiety disorders, the evidence with regard to severity and antidepressant efficacy for anxiety is very limited, although some small studies have suggested that there is no association between initial severity and

antidepressant efficacy for anxiety disorders [77, 78]. Individual participant data meta-analyses have also found no evidence that initial severity of MDD moderates the efficacy of CBT compared to pill placebo [79] or antidepressants [80], in contrast to a widely-held belief that psychotherapy may be sufficient for mild depression, while antidepressants are necessary for severe depression.

Relatively few studies have examined whether clinical predictors could inform treatment selection [81]. Two studies have used data from the large STAR*D (Sequenced Treatment Alternatives to Relieve Depression) trial to predict who will (or will not) respond to antidepressant treatment using baseline clinical and demographic information [82, 83]. One of these studies [83] also found that the model appeared to be specific to treatment with citalopram and did not predict response to combined venlafaxine and bupropion, which suggests that it could be used for initial treatment selection. However, further validation and improvement would be necessary, as accuracy was still rather low (60%).

Studies in small samples have also suggested that simple clinical predictors could identify the optimal treatment for specific patients, when comparing antidepressants and CBT [84], or interpersonal psychotherapy and cognitive therapy [85]. The continued development of statistical learning techniques and the increased accessibility of large sample sizes through individual participant data meta-analysis are likely to offer opportunities to further improve these models.

However, at present, the possibilities for determining who will benefit from (which) treatment before the start of treatment are still limited. Another field of research, therefore, is aimed at investigating whether it is possible to determine whether a patient will benefit from treatment earlier in the course of treatment than is currently done. For instance, current guidelines usually recommend at least 4 and sometimes up to 8 weeks of treatment with an adequate dose of an antidepressant before considering switching antidepressants or augmenting treatment with other medications or psychotherapy [73, 74, 86]. However, many patients show a detectable, if modest, improvement within the first two weeks of treatment, and this early improvement has been robustly associated with attaining a full response or remission by the end of treatment [87, 88, 89, 90, 91, 92, 93].

While some of these studies have suggested that lack of early improvement is a sufficiently good predictor that a change in management is indicated for patients who do not show any improvement by two weeks [87], others have found that these patients still had a reasonably good chance of responding later in the study, which suggests that it would be premature to switch or augment treatment [94, 95]. Better predictive models, therefore, are desirable.

Recent research has suggested that symptoms are not interchangeable and the sum score on a depression or anxiety questionnaire may conceal important information [96]. For instance, symptoms such as sad mood or concentration problems appear to be more strongly associated with functional impairments than insomnia or changes in appetite

[97]. Improvement in specific symptoms is associated with a good response [98, 99, 100], but no study so far has controlled for the improvement in the sum score, so it is unclear whether examining individual symptoms can help to enhance predictive models.

Thesis outline

In this thesis, I aimed to bring the evidence base for treating depression and anxiety to light, particularly where it concerns antidepressants. In Part I, chapters 2 through 8, I study the impact of reporting and citation biases on the evidence base and also investigate whether the best available evidence (as synthesized in clinical guidelines) is actually put into practice. In Part II, chapters 9 through 11, I examine whether routine clinical information, specifically initial severity and early improvement in individual symptoms, can be used to predict who will benefit from antidepressants.

Part I: Bringing the evidence to light

Chapter 2 investigates reporting bias in clinical trials of antidepressants for the short-term treatment of anxiety disorders, focusing on the primary efficacy outcome. Using Food and Drug Administration (FDA) reviews, a complete cohort of trials was assembled and traced into the published literature to determine their fate. **Chapter 3** builds upon the results in chapter 2 by examining the oft-neglected flip side of the coin, namely safety. **Chapter 4** explores whether the practice of pooling trials for publication constitutes an additional type of reporting bias.

Chapters 5 and 6 examine spin (or positive focus) and citation bias. For this, I use the highly controversial literature on 5-HTTLPR, which has also been suggested to play a role in antidepressant response. I specifically look at a) gene-environment interactions in the development of depression and b) amygdala activation as an underlying mechanism for the development of depression. I examine whether spin and citation bias could play a role in the persistence of belief in spite of an unreliable evidence base.

The different reporting and citation biases are often examined separately, but in **Chapter 7** I examine the pernicious cumulative impact of study publication bias, selective outcome reporting bias, spin, and citation bias on the apparent efficacy of antidepressants and psychotherapy for depression.

This part concludes with **Chapter 8**, in which I study to what extent the evidence is actually put into practice. A prescription database is used to examine whether physicians adhere to the guidelines for antidepressant initiation in children and adolescents.

Part II: Who benefits from antidepressants?

Chapters 9 and 10 investigate the possible influence of initial severity on antidepressant efficacy for anxiety disorders, following the high-profile (but controversial) findings that antidepressants have minimal efficacy in mild depression. In **Chapter 9**, I look at whether the average baseline severity in a trial is associated with the antidepressant-placebo difference. As there are disadvantages to using trial averages, I study this question in more detail using individual participant data in **Chapter 10**.

Finally, in **Chapter 11**, I use early improvement to predict which depressed patients will benefit from antidepressants. In particular, I examine the predictive value of improvement in individual depressive symptoms, to determine whether examining individual symptoms can result in better predictive models than examining the total score alone.

Part I

Bringing the evidence to light

Chapter 2

Reporting bias in clinical trials investigating the efficacy of second-generation antidepressants in the treatment of anxiety disorders

Annelieke M. Roest, Peter de Jonge, Craig D. Williams,
Ymkje Anna de Vries, Robert A. Schoevers, Erick H. Turner

JAMA Psychiatry (2015), 72 (5), 500 - 510

Abstract

Importance: Previous studies have shown that the scientific literature has overestimated antidepressant efficacy for depression, but other indications have not been considered.

Objective: To examine reporting biases in clinical trials of antidepressants for anxiety disorders, and to quantify the extent to which these biases inflate estimates of drug efficacy.

Data sources and study selection: Reviews of premarketing trials for 9 second-generation antidepressants were obtained from the Food and Drug Administration (FDA). A systematic search for matching publications was performed using PubMed, EMBASE and Cochrane CENTRAL.

Data extraction and synthesis: Double data extraction was performed for the FDA reviews and the journal articles. Hedges' g was calculated as measure of effect size.

Main outcomes and measures: Reporting bias was classified as study publication bias, outcome reporting bias, or spin. Separate meta-analyses were conducted for the two sources and meta-regression was used to assess the impact of publication status on effect estimates.

Results: Sixteen of 57 (28%) trials were not positive according to the FDA, while only 2 of 45 (4%) published article conclusions were not positive ($p<0.001$). Positive trials were 5 times more likely to be published in agreement with the FDA determination compared to trials determined not-positive (risk ratio=5.20; 95% CI: 1.87 - 14.45; $p<0.001$). We found evidence for study publication bias ($p<0.001$), outcome reporting bias ($p=0.02$), and spin ($p=0.02$). The pooled effect size based on the published literature ($g=0.38$; 95% CI: 0.33 - 0.42; $p<0.001$) was 15% higher than the effect size based on the FDA data ($g=0.33$; 95% CI: 0.29 - 0.38; $p<0.001$), but this difference was not statistically significant ($p=0.18$).

Conclusions and relevance: Various reporting biases were present for trials on the efficacy of FDA-approved second-generation antidepressants for anxiety disorders. Although this did not significantly inflate estimates of drug efficacy, reporting biases led to significant increases in the number of positive findings in the literature.

Introduction

There is strong evidence that significant results from randomized controlled trials are more likely to be published than nonsignificant results [38]. As a consequence, the published literature, including meta-analyses, may overestimate the benefits of treatment while underestimating its harms, thus misinforming clinicians, policy makers, and patients [101].

Different types of reporting biases can be present. Study publication bias occurs when studies with positive results are more likely to be published than studies with negative results [102]. Outcome reporting bias involves publishing outcomes from a study that are "positive" (e.g., statistically significant) without publishing "negative" outcomes or switching the status of primary and secondary outcomes based on results [103]. Finally, spin occurs when treatments are described by investigators as beneficial, even though published results for primary outcomes are nonsignificant [40].

The registry and results database of the Food and Drug Administration (FDA) can be used to assess the degree to which published trial results may overestimate efficacy [19, 104, 105, 106]. Pharmaceutical companies must register all trials they intend to use in support of an application for US marketing approval with the FDA, and information on these trials is compiled in this database. A previous study found that 51% of trials of antidepressants for major depressive disorder were deemed positive by the FDA compared to 94% in the published literature; in addition, a meta-analysis of only published data overestimated the effect of antidepressants by 32% [19]. This was followed by debate and additional research on the efficacy of antidepressants for depression [72, 101, 107, 108].

Antidepressants are widely prescribed for conditions other than depression [109], including anxiety disorders. However, research on reporting biases for these other indications is lacking. Anxiety disorders are common in the general population with an estimated year prevalence of 12% [110]. Second-generation antidepressant drugs, namely selective serotonin reuptake inhibitors (SSRIs) and serotonin norepinephrine reuptake inhibitors (SNRIs), are the primary pharmacological treatments for generalized anxiety disorder (GAD) [13, 111], panic disorder (PD) [13, 112], social anxiety disorder (SAD) [113], post-traumatic stress disorder (PTSD) [114], and obsessive compulsive disorder (OCD) [115].

Several meta-analyses have reported that second-generation antidepressants are superior to placebo in the treatment of GAD [116, 117], PD [118, 119], SAD [120, 121], PTSD [122] and OCD [123]. Some of these meta-analyses suggested the existence of study publication bias based on funnel plot asymmetry [118, 120]. However, such methods cannot prove the existence of publication bias; for that, one must access and analyze unpublished data as well [106]. A recent study examined the efficacy of one SSRI in the treatment of GAD and PD using a complete dataset of trials sponsored by the manufacturer. This study indeed showed that published trials had significantly larger effect sizes than unpublished trials [77].

In the present study, the first objective was to examine reporting bias in the scientific literature on efficacy of second-generation antidepressants that are FDA-approved for the treatment of anxiety disorders. By comparing published articles with the corresponding FDA reviews we examined the presence of study publication bias, outcome reporting bias, and spin. The second objective was to compare the magnitude of the overall effect based on published trial data from premarketing trials with that based on the full cohort of such trials registered with the FDA.

Methods

As in previous studies [19, 104], we began by identifying the inception cohort of premarketing trials for the indications of interest, then conducted a literature search for those trials.

Data from FDA reviews

We identified the phase 2/3 clinical double-blind placebo-controlled trials registered with the FDA and conducted in pursuit of marketing approval of second-generation antidepressants for the treatment of the following five disorders: (1) GAD, (2) PD, (3) SAD, (4) PTSD, and (5) OCD. Nine drugs, approved by the FDA for these indications, were examined: seven SSRIs (paroxetine, paroxetine controlled release [CR], sertraline, fluoxetine, fluvoxamine, fluvoxamine CR, and escitalopram) and two SNRIs (venlafaxine extended release [ER] and duloxetine). We retrieved the FDA Drug Approval Packages (aka FDA reviews) from the FDA's website; if these were not available for download, we requested them from the FDA's Freedom of Information Office (http://www.accessdata.fda.gov/scripts/foi/FOIRequest/requestinfo.cfm).

We extracted the results the FDA used to decide whether the trial was positive, i.e. whether it could be used to support marketing approval. Data were extracted preferably from the statistical review, but also from the medical review and administrative correspondence (e.g. memos by team leader). In cases where multiple primary endpoints were identified in a trial, results were extracted for the endpoint that was most consistent with the primary endpoint identified in other trials for the same indication.

In accordance with previous publications [19, 104], the FDA's regulatory decisions were classified as (1) positive (clearly supporting efficacy), or (2) not positive, with the latter including both questionable (neither clearly positive nor clearly negative) and negative trials (not supportive of efficacy).

The questionable category included trials characterized by the FDA as "marginally" or "borderline" positive. These were trials that had non-significant p values for one or more

of the primary endpoints, but were considered by the FDA to be supportive of other positive trials because of significant findings on secondary variables. The questionable category also included "failed" trials (in which neither the study drug nor the active comparator demonstrated statistical superiority to placebo).

For multiple-dose trials, we used the FDA's overall decision on the trial. For purposes of meta-analysis, we extracted data only for approved dosages, thus excluding "subtherapeutic" dosages [19]. Data extraction, classification, and data entry was performed independently by two investigators (AR and CW) with discrepancies resolved by consensus (AR, CW, ET).

Data from journal articles

Having identified the inception cohort of premarketing trials registered with the FDA, we systematically searched for matching publications using PubMed, EMBASE and the Cochrane Central Register of Controlled Trials (CENTRAL) without language restrictions, with a search cutoff date of December 19, 2012. We searched the title field for the name of the drug and the type of anxiety disorder, and any field for the word "placebo". For example, when searching PubMed for relevant escitalopram trials for GAD, the search syntax was "escitalopram[Title] AND (generalized[Title] OR generalised[Title]) AND anxiety[Title] AND disorder[Title] AND placebo."

Publication matches for trials registered with the FDA were identified using the following information: drug name, name of the active comparator (if applicable), dosage groups, sample sizes, trial duration, and names of investigators. The preferred type of publication was a stand-alone publication, i.e. a full-length article devoted to reporting the results of a single trial. If no stand-alone publication could be found, then pooled analyses were sought in which multiple trials were covered in a single article. Data from journal articles that pooled data from multiple trials that were not identical in design according to the FDA were excluded from this study. Pooled-trials publications were also excluded when one or more of the included trials were published earlier as stand-alone publications and the pooled-trials publication did not present separate results for the included trials. Finally, data published only in abstract form were excluded.

Several steps were taken to minimize the possibility that we missed matching publications. If no publication was found via the electronic database search, PubMed was used to identify the three most recent review articles focusing on the efficacy of the trial drug for the condition treated in the trial. The reference lists for those publications were hand searched. In addition, the drug sponsor's website was searched for bibliographic information on the trials in question.

To assess drug efficacy according to published journal articles, we used the primary endpoint specified in the publication. If a primary endpoint was not specified and if no

endpoint was clearly emphasized, we extracted the drug-placebo comparison reported first in the text of the results section or in the table or figure first cited in the text [104]. If multiple endpoints were identified as primary in a single study, results were extracted for the endpoint reviewed as primary by the FDA. Data extraction and entry was done independently by AR and RS with discrepancies resolved through consensus (AR, RS, YV).

In addition, each article's conclusion was classified as positive or not-positive (including questionable and negative) based on the sentence in the abstract reporting the authors' overall conclusion regarding study outcome. Conclusions were classified independently by AR and PJ, who was blinded to the results of the FDA review.

Statistical analysis

All statistical analyses were performed using STATA 11.0. The binomial probability test was used to assess whether the proportion of positive conclusions in journal articles was significantly different from the proportion of positive trials according to the FDA. In addition, we examined whether not-positive trials (according to the FDA) were more likely to be unpublished, or published in a positive manner, compared with positive trials using Fisher's exact test. The presence of study publication bias (trial results not published), outcome reporting bias (changes in analysis or primary endpoint affecting significance of findings), and spin (abstract conclusion not consistent with published results on primary endpoint) was also compared for positive and not-positive trials.

We conducted two meta-analyses: one using data from the FDA reviews and another one using the corresponding published data [19, 104]. Hedges' g was used as the measure of effect size and was calculated using the following equation in which t represents the t statistic and n_1 and n_2 are the numbers of subjects in the drug and placebo groups, respectively:

$$g = t \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

The values for g were adjusted using Hedges' correction for small sample size [19, 104]. The t statistic was calculated from the precise p value and the trial sample size using Microsoft Excel's TINV function, multiplying t by -1 when the study drug was inferior to placebo. If a precise p value was not available because it was reported as a range (e.g. $p<0.05$), the t statistic was calculated from other summary statistics, namely standard deviations, standard errors, and 95% confidence intervals around the mean difference. When the data were presented as dichotomized statistics, Hedges' g was calculated from χ^2 [122]. If none of these data were available, and FDA and journal data were otherwise congruent, data were imputed with data extracted from the other source. Additionally,

for two journal articles Hedges' g was calculated from the F statistic (analysis of variance) [124]. Finally, p values and other efficacy data were not reported for 2 negative FDA trials that were, in one case, not published and, in the other case, published as positive. These p values were imputed with p=0.396, which was derived from 16 nonsignificant but precise P values, according to the method previously described by Turner et al. [19].

For each multiple-dose study, we computed a single study-level effect size using a fixed effects model to pool the values from that trial's multiple treatment arms. When calculating the standard error, each trial's shared placebo n was counted once, rather than redundantly, for each dose group to avoid a spuriously low standard error. A limitation of this method is that it only partially addresses error due to correlation between the comparisons [102]. Calculations of all effect sizes were performed independently by AR and YV.

The random effects pooling method was used to generate summary estimates of Hedges' g. I^2 and confidence intervals around I^2 were calculated to assess heterogeneity [125]. I^2 reflects the proportion of total variance explained by heterogeneity. Meta-regression, using the restricted maximum likelihood method, was conducted to examine the impact of publication status on the effect estimates. In addition, pre-specified subgroup analyses were performed for each anxiety disorder.

Results

FDA reviews

We analyzed 9 second-generation antidepressants for data related to the 5 anxiety disorders. Within those 45 possible drug-indication combinations, 21 are FDA-approved. Of those, we were able to download 9 FDA approval packages through the FDA website; for the remaining 12, we made requests to the FDA Freedom of Information Office. Of these, the FDA Freedom of Information Office fulfilled 11 requests — the FDA informed us that the drug approval package for fluoxetine for panic disorder would not be available for at least 18 months. This left 20 approval packages in this study. These drug approval packages, which were issued between 1994 and 2008, reviewed the results of 57 randomized, placebo-controlled short-term trials.

Journal articles

For the 57 above-mentioned FDA-registered trials, we identified 52 trials published in 48 publications. Three of these articles were excluded from further analyses. As a result, 3 additional trials were judged to be not fully published. Two articles pooled trials which were not identical in design [126, 127] and another pooled-trials article failed to present

separate results for the included trials [128] and included a trial that was previously published as a stand-alone publication [129].

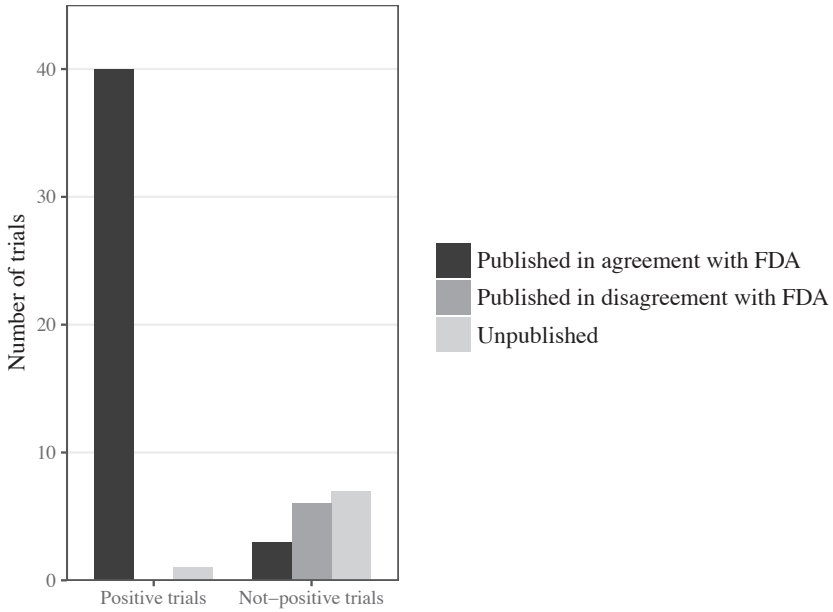

Figure 2.1: *Publication status of positive and not-positive FDA trials*

Trial outcome versus published results

The proportion of positive findings was 72% (41/57) according to the FDA versus 96% (43/45) according to the published literature. This difference was statistically significant (binomial test p<0.001).

Of the 41 positive trials, 40 (98%) were published in agreement with the FDA (Figure 2.1). By contrast, of the 16 not-positive trials, only 3 (19%) were published in agreement with the FDA. This difference was statistically significant (Fisher's exact test p<0.001). Overall, trials that the FDA judged as positive were 5 times more likely to be published in agreement than FDA-not-positive trials (risk ratio: 5.20; 95% CI: 1.87 - 14.45; p<0.001).

Study publication bias

Seven of the 16 (44%) not-positive trials were not published, while only 1 of the 41 (2%) positive trials was not published (Table 2.1). This difference was statistically significant (Fisher's exact test p<0.001).

Table 2.1: *Characteristics of included premarketing trials*

			N				
Disorder	Drug	Trial	Placebo	Drug	Outcome	FDA	Bias
GAD	Escitalopram	MD-05 [130]	128	124	HAM-A	P	-
		MD-06 [130]	138	143	HAM-A	P	-
		MD-07 [131]	153	154	HAM-A	P	-
	Paroxetine	641 [132]	180	385	HAM-A	P	-
		642 [133]	163	161	HAM-A	P	-
		637 [N/A]	183	181	HAM-A	N	SPB
	Duloxetine	HMBR [134]	173	334	HAM-A	P	-
		HMDT [135]	158	161	HAM-A	P	-
		HMDU [136]	158	149	HAM-A	P	-
	Venlafaxine ER	210 [137]	96	253	HAM-A	P	-
		214 [138]	98	174	HAM-A	P	-
PD	Paroxetine	120 [139]	69	72	% 0 attacks	Q	ORB
		108 [140]	60	60	% 0 or 1 attack	P	-
		187 [141]	123	123	% 0 attacks	P	-
		223 [N/A]	68	77	% 0 attacks	Q	SPB
	Paroxetine CR	494 [142]	129	122	% 0 attacks	P	-
		495 [142]	136	123	% 0 attacks	P	-
		497 [142]	130	132	% 0 attacks	N	-
	Sertraline	629 [143]	87	79	# of attacks	P	-
		630 [144]	88	88	# of attacks	P	-
		529 [129]	44	127	# of attacks	Q	ORB
		514 [N/A]	38	112	# of attacks	N	SPB
	Venlafaxine ER	398 [145]	154	315	% 0 attacks	P	-
		399 [146]	157	316	% 0 attacks	P	-
		353 [147]	155	155	% 0 attacks	Q	-
		391 [148]	168	160	% 0 attacks	N	Spin
SAD	Fluvoxamine CR	3107 [149]	125	110	LSAS	P	-
		3108 [150]	148	126	LSAS	P	-
	Paroxetine	502 [151]	145	136	LSAS	P	-
		382 [152]	92	90	LSAS	P	-
		454 [153]	92	268	LSAS	P	-
	Paroxetine CR	790 [154]	184	185	LSAS	P	-
	Sertraline	R-0601 [155]	196	205	LSAS	P	-
		94-004 [156]	69	134	BSPS	P	-
		95-003 [157]	196	191	CGI-L	N	ORB
	Venlafaxine ER	387 [158]	138	133	LSAS	P	-
		393 [159]	135	126	LSAS	P	-
PTSD	Sertraline	641 [160]	82	84	CAPS-2	N	-
		682 [N/A]	94	94	CAPS-2	N	SPB
		640 [161]	104	98	CAPS-2	P	-
		671 [162]	90	93	CAPS-2	P	-
	Paroxetine	651 [163]	167	322	CAPS-2	P	-

continued

Table 2.1: *Characteristics of included premarketing trials*

Disorder	Drug	Trial	N		Outcome	FDA	Bias
			Placebo	Drug			
		648 [164]	133	136	CAPS-2	P	-
		627 [N/A]	159	154	CAPS-2	Q	SPB
OCD	Fluoxetine	HCEP 1 [165]	47	139	Y-BOCS	P	-
		HCEP 2 [165]	41	122	Y-BOCS	P	-
		E079 [166]	56	158	Y-BOCS	N	Spin
	Fluvoxamine	5529 [N/A]	80	79	Y-BOCS	P	-
		5534 [167]	77	78	Y-BOCS	P	-
	Fluvoxamine CR	3103 [168]	119	113	Y-BOCS	P	-
	Paroxetine	116 [169]	88	166	Y-BOCS	P	-
		118 [N/A]	75	79	Y-BOCS	N	SPB
		136 [170]	99	198	Y-BOCS	P	-
	Sertraline	237/248 [171]	44	43	Y-BOCS	Q	Spin
		371/372 [172]	84	240	Y-BOCS	P	-
		546 [173]	79	85	Y-BOCS	P	-
		495 [N/A]	87	83	Y-BOCS	N	SPB

The FDA column indicates the Food and Drug Administration (FDA) decision (P: positive; N: negative; Q: questionable). Type of bias includes study publication bias (SPB), outcome reporting bias (ORB), and spin; "-" indicates that no bias was present. Other acronyms – BSPS: Brief Social Phobia Scale; CAPS-2: Clinician-Administered PTSD Scale, part 2; CGI-L: Clinical Global Impressions-Liebowitz; GAD: generalized anxiety disorder; HAM-A: Hamilton Rating Scale for Anxiety; LSAS: Liebowitz Social Anxiety Scale; OCD: obsessive compulsive disorder; PD: panic disorder; PTSD: posttraumatic stress disorder; SAD: social anxiety disorder; Y-BOCS: Yale-Brown Obsessive Compulsive Scale

Outcome reporting bias

For 3 of the 16 not-positive trials (19%), results were published with a conclusion that conflicted with that in the FDA review, changing their effects from nonsignificant to statistically significant. By contrast, outcome reporting bias was found in none of the 41 FDA-positive trials (Table 2.1). The difference in proportions was statistically significant (Fisher's exact test p=0.02).

One of the 3 above-mentioned publications (trial 120, paroxetine for PD) presented only observed-cases analyses for the primary outcome [139]; according to the FDA, the primary analysis involved last-observation-carried-forward (LOCF) analyses, the results of which were not statistically significant.

In the article presenting results of trial 529, data from subjects with PD who were randomized to different dosages of sertraline were pooled and compared to the placebo group, yielding a significant result [129]; the FDA review showed that the primary results for each of the individual dosage groups were nonsignificant.

Finally, one article presenting the results of trial 95-003, which compared the effect of sertraline (with and without exposure therapy) to placebo (with and without exposure therapy) in patients with SAD, combined scores on three endpoints (disorder-specific Clinical Global Impression Scale [severity and improvement] and Social Phobia Scale) in response versus non-response categories [157]; the FDA review showed that the primary endpoint was the severity total score of the disorder-specific Clinical Global Impression Scale and that this was nonsignificant.

Spin

Spin was present in an additional 3 out of 16 (19%) of the not-positive trials and not present for positive trials (Fisher's exact test p=0.02) (Table 2.1). Each of these 3 articles [166, 171, 148] reported that the primary endpoint was nonsignificant in the results section but, in the abstract, concluded that the trial was positive. The FDA classified these trials as questionable (trial 237/248: sertraline for OCD) or negative (trial E079: fluoxetine for OCD and trial 391: venlafaxine ER for PD). Conclusions on study drug efficacy for these trials, according to the FDA and the authors of the journal articles, are included in Table 2.3 in the Appendix.

Meta-analysis

The pooled effect size based on the FDA data was 0.33 (95% CI: 0.29 - 0.38; p<0.001). Heterogeneity was moderate (I^2=39%; 95% CI: 15% - 56%). For trials published in agreement with the FDA review results, the pooled effect size (Hedges' g=0.38; 95% CI: 0.34-0.42; p<0.001) was larger than the pooled effect size of trials that were not published or published in disagreement with the FDA conclusion (g=0.17; 95% CI: 0.09 - 0.26; p<0.001). Meta-regression showed this difference to be statistically significant (g=0.21; 95% CI: 0.12 - 0.30; t=4.61; p<0.001).

The pooled effect size based on the published literature was 0.38 (95% CI: 0.33 - 0.42; p<0.001). Heterogeneity was low (I^2=30%; 95% CI: 0% - 51%). This effect size represented a 15% increase in effect size compared with the value based on the FDA data. This difference was not statistically significant by meta-regression (g=0.04; 95% CI: -0.02 - 0.10; t=1.36; p=0.18).

Effect sizes based on data from the FDA reviews were 0.32 for GAD, 0.28 for PD, 0.27 for PTSD, and 0.39 for both OCD and SAD. For all disorders the pooled effect sizes of trials published in agreement with the FDA review results were larger than the pooled effect sizes of trials that were not published or published in disagreement with the FDA conclusion (Table 2.2). As a result forest plots for all disorders showed fewer nonsignificant trials according to the published literature than according to the FDA, especially for PD

Table 2.2: *Meta-analysis*

	FDA					
	Total	Published in agreement	Not published in agreement	Journal	Overestimate	
Disorder	Hedges' g (95% CI)	Hedges' g (95% CI)	Hedges' g (95% CI)	Hedges' g (95% CI)	%	*p*
GAD	0.32 (0.25-0.39)	0.34 (0.28-0.41)	0.11 (-0.09-0.31)	0.34 (0.27-0.41)	6%	0.65
PD	0.28 (0.20-0.36)	0.33 (0.25-0.41)	0.13 (-0.02-0.29)	0.35 (0.24-0.46)	25%	0.38
SAD	0.39 (0.30-0.49)	0.43 (0.35-0.50)	0.09 (-0.11-0.29)	0.42 (0.35-0.49)	8%	0.56
PTSD	0.27 (0.11-0.44)	0.33 (0.14-0.53)	0.13 (-0.12-0.38)	0.32 (0.14-0.50)	19%	0.76
OCD	0.39 (0.30-0.49)	0.44 (0.33-0.54)	0.30 (0.11-0.48)	0.45 (0.35-0.56)	15%	0.42
Overall	0.33 (0.29-0.38)	0.38 (0.34-0.42)	0.17 (0.09-0.26)	0.38 (0.33-0.42)	15%	0.18

FDA: Food and Drug Administration; GAD: generalized anxiety disorder; OCD: obsessive compulsive disorder; PD: panic disorder; PTSD: posttraumatic stress disorder; SAD: social anxiety disorder.

and OCD (Figures 2.2 and 2.3 [see Figures 2.4, 2.5, and 2.6 in the Appendix for GAD, SAD, and PTSD]).

Effect sizes based on the literature were larger for all disorders as compared with effect sizes based on the FDA reviews, with the smallest increases for GAD (g=0.34, 6% increase) and SAD (g=0.42, 8% increase) and larger increases for OCD (g=0.45, 15% increase), PTSD (g=0.32, 19% increase) and PD (g=0.35, 25% increase). However, the differences in effect estimates based on the journal articles and the FDA reviews were not statistically significant for any of the individual disorders (Table 2.2).

Discussion

This study showed the presence of reporting bias in randomized controlled trials on the efficacy of second-generation antidepressants for anxiety disorders. Trials that the FDA judged to be positive were over 5 times more likely to be published in agreement with the FDA analysis than not-positive trials. As a result, 96% (43/45) of the journal articles were framed positively, while 72% (41/57) of the trials were deemed positive by the FDA. All examined reporting biases were present among the included trials, namely study publication bias, outcome reporting bias, and spin.

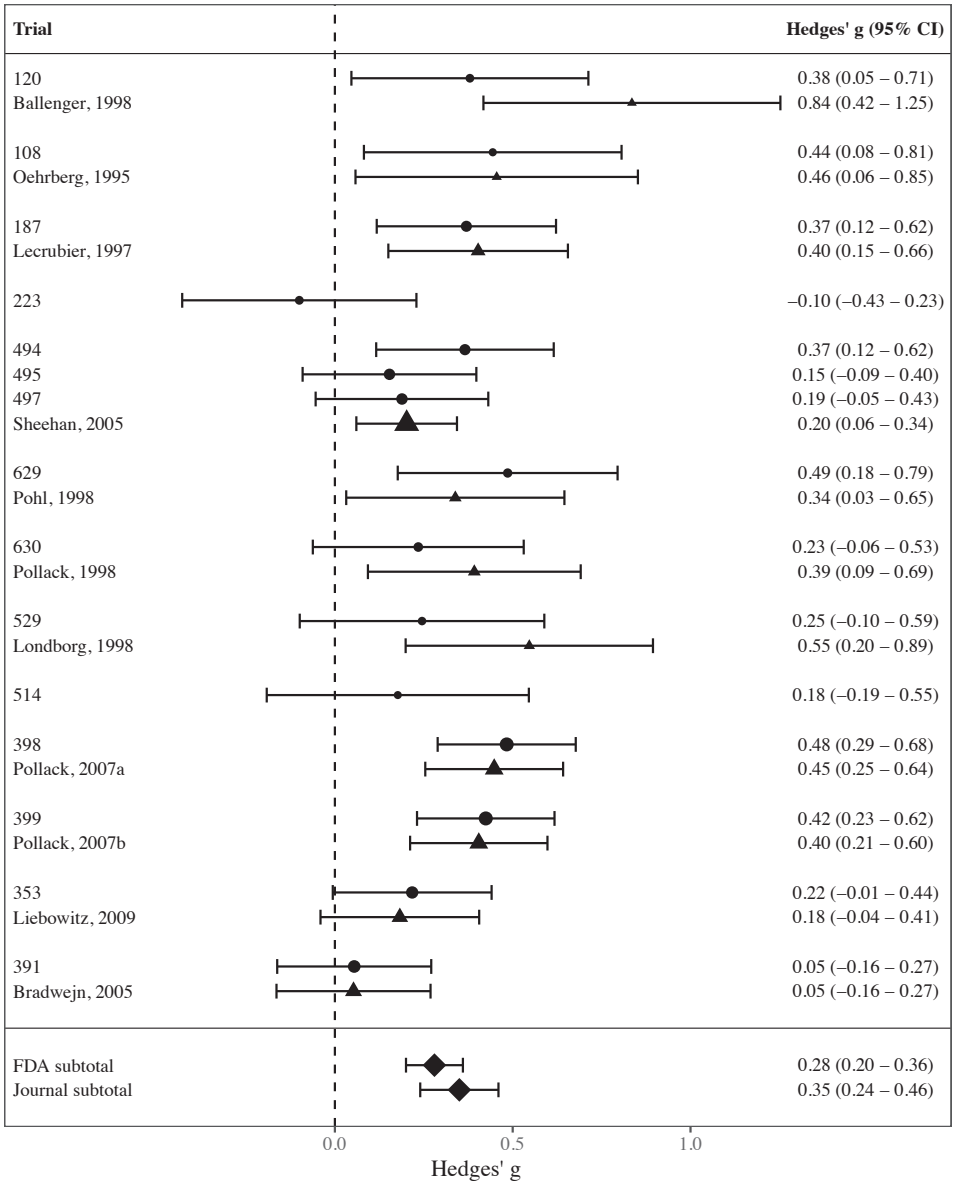

Figure 2.2: *Forest plot for PD. Circles indicate FDA effect sizes, while triangles indicate the matching journal effect sizes (for published trials).*

In a previous study that examined reporting bias in trials on second-generation antidepressants for major depressive disorder [19], the overall effect size based on the FDA data was 0.31, quite comparable to the effect size of 0.33 found in this study. After conducting two meta-analyses, one based on data from the FDA reviews and the other based on data from the corresponding journal articles, we found that reporting bias inflated the

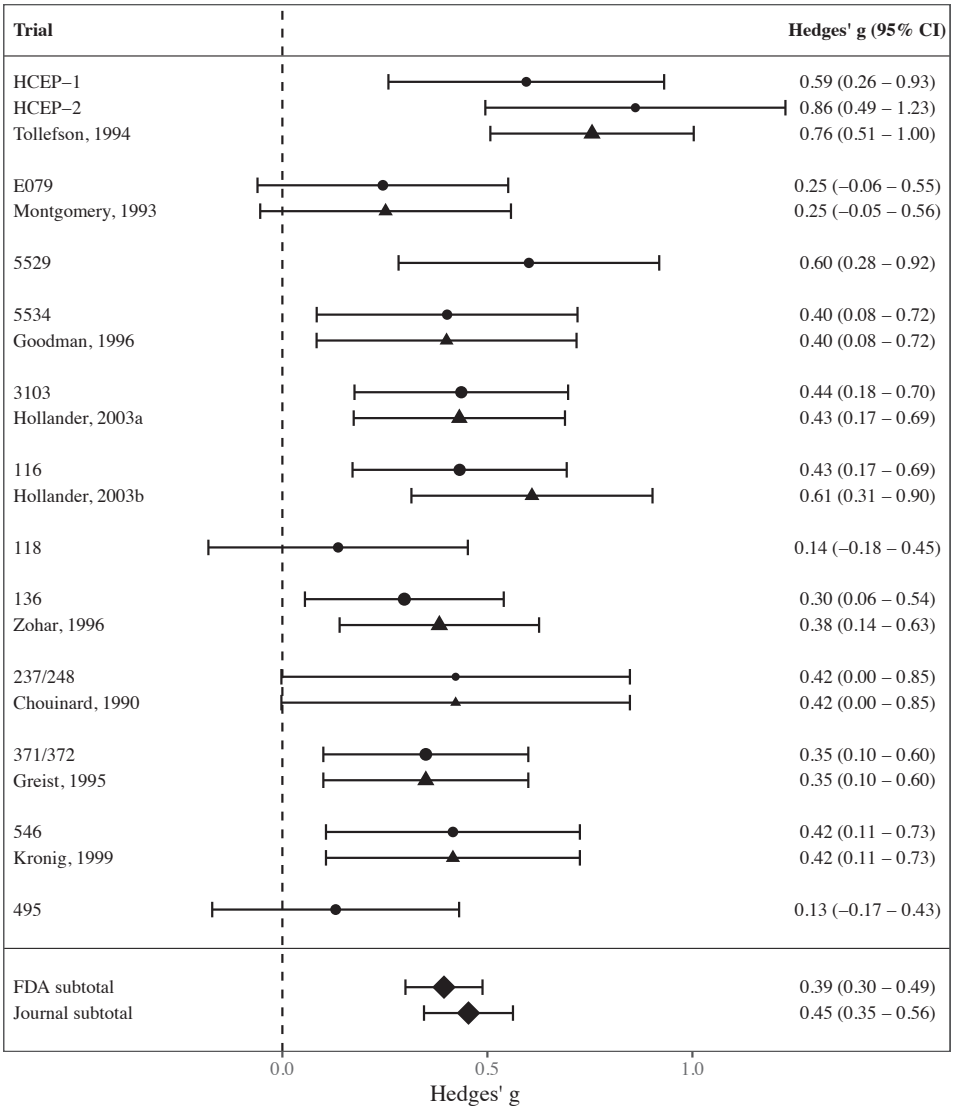

Figure 2.3: *Forest plot for OCD. Circles indicate FDA effect sizes, while triangles indicate the matching journal effect sizes (for published trials).*

apparent effect size by 15%. This increase was not statistically significant, in contrast to the larger inflation factor (32%) found earlier with major depressive disorder [19]. For the individual anxiety disorders, the inflation factors ranged from 6% (GAD) to 25% (PD), indicating the importance of using unbiased data in meta-analyses on the efficacy of second-generation antidepressants for anxiety disorders.

In the main analyses we combined five disorders classified as anxiety disorders in DSM-

IV; however, in DSM-V, OCD is now classified under obsessive-compulsive and related disorders, while PTSD is under trauma- and stressor-related disorders. Therefore, with the recent change in taxonomy, our grouping of these disorders could be viewed as a limitation, although efficacy of the drugs was comparable across disorders.

A clearer limitation of the current study is that the trials for the individual anxiety disorders were few in number, decreasing the power of the subgroup analyses. An additional limitation is that we did not examine biased reporting of harm outcomes, which figures into the overall risk-benefit ratio of a drug, but such an examination would have been beyond the scope of the current study.

Certain data available only in pooled-trials publications were classified as unpublished, which could also be viewed as a limitation. However, a study of antidepressant trials submitted to the Swedish drug regulatory authority showed that positive trials were more likely to be published as stand-alone publications, while negative trials tended to be reported only within pooled-trials publications [27]. Pooled analyses may not follow the predetermined analysis plan and power calculation and can therefore yield different conclusions than the original trials. Pooled analyses are also associated with "salami slicing" (publishing similar results from one study in multiple publications) [174]. Therefore, although these publications may provide new information, especially on subgroups and secondary endpoints, they are susceptible to bias [175]. This bias can be reduced by first publishing the original trial results. Future research could assess the bias that is introduced by pooled-trials publications.

Finally, we did not contact drug sponsors to ask whether specific trials were published in the scientific literature, so there is a small chance that trials could have been misclassified as unpublished. However, considering the extensive literature search methods, it seems unlikely that such trial publications would be discoverable by the typical health care professional.

A strength of this study is that, for 20 of the 21 FDA-approved drug-indication combinations, we were able to include data from all premarketing randomized controlled trials, thereby allowing a reliable assessment of different reporting biases for these trials. However, it is important to note that we could not include data from rejected drug-indication applications because the FDA does not release these reviews [105]. It is likely that the amount of reporting bias that was found would increase if these trials were to be examined as well.

In addition, our estimation of the amount of reporting bias present might also be influenced by the fact that all trials were sponsored by pharmaceutical industries. Yet reporting bias is not restricted to pharmacological treatments sponsored by drug companies [176]. Since reporting bias has been shown for the treatment of depression with psychotherapy [177], it should be worthwhile to systematically assess reporting bias in trials using psychotherapy for anxiety disorders as well.

Spin can result from different, intentional or unintentional, strategies, for example by focusing on secondary endpoints for which significant results were obtained [178]. Journal articles for which spin was identified also often reported "marginally significant results" (p values between 0.05 and 0.10) in the present study. Ideally, interpretation of trial results should not be based solely on a p value indicating whether results are statistically significant or not [40]. In addition to providing p values, future research could consider including Bayes factors as a measure of the strength of the evidence. Bayes factors stem from Bayesian statistics and have the advantage that they can express the strength of the evidence on a continuous scale [179].

Reporting bias significantly increased the number of positive versus negative publications in the literature in the present study. This likely impacts clinician's perceptions of the efficacy of these drugs, which could reasonably be expected to affect prescription behavior. In both Europe and in the US, use of antidepressants has been rising markedly over the last two decades with much of that use appearing to be driven by non-specialists in settings of primary care [180, 181]. Although it should be noted that these studies could not take into account indications for which the drugs were prescribed, a realistic view of the efficacy of these agents is important across all indications. Results of the current study and other studies comparing published results to data from FDA reviews or other registries [19, 77] can perhaps assist clinicians in gaining a more realistic view of the evidence for the efficacy of antidepressants in the (short-term) treatment of affective disorders.

This study adds to the growing body of literature establishing the pervasiveness of reporting bias [34, 176, 182]. It also highlights the need to address this problem using various measures, as recently reviewed [183]. One suggested approach, which would address outcome reporting bias and spin (but not study publication bias), would require peer reviewers to make preliminary decisions based on the strength of the methods in the original trial protocol [176] so that their decisions are not influenced by the statistical significance of study results [184]. Use of study registries, like ClinicalTrials.gov, can also reduce reporting bias in the scientific literature [176], but this registry does not yet function optimally. For example, for the majority of trials subjected to mandatory reporting within one year following trial completion, results were not posted within this timeframe [185].

In summary, although the majority of trials on the efficacy of FDA-approved second-generation antidepressants for anxiety disorders were positive, various reporting biases were present. These reporting biases led to an overly positive representation of significant findings in the scientific literature.

Appendix

Table 2.3: *FDA conclusion vs. journal conclusion for articles with spin*

Drug	Disorder	Trial	FDA	Literature
Venlafaxine ER	PD	391	"The results of study [391]* do not provide adequate evidence of the anti-panic efficacy of venlafaxine ER versus placebo over 10 weeks of treatment"	"Venlafaxine ER seems to be effective and well tolerated in the short-term treatment of PD"
Fluoxetine	OCD	E079	"This trial failed to show significant fluoxetine-placebo differences on any of the scales. A few contrasts were marginal, with p-values between 0.05 and 0.10, but considering the number of scales and number of tests done, I attach no importance to them."	"This study supports the growing evidence for the safety and efficacy of fluoxetine in the treatment of OCD."
Sertraline	OCD	237/ 248	"Although the plots of group means over time appear to show that sertraline beats placebo, differences between group means were inconsistent among the various time points, with significance appearing somewhat early in the study and frequently disappearing towards week 8. It would be difficult to characterize these results as positive, and at best we might call them supportive. Calling it a failed study may be more accurate."	"Results of the Y-BOCS total score, the NIMH score, and the global severity and improvement scores demonstrated a statistically significant superiority of sertraline compared with placebo."

*ER: extended release; FDA: Food and Drug Administration; OCD: obsessive compulsive disorder; PD: panic disorder. FDA conclusions were extracted from the medical review (trial 391) or the statistical review (trials E079 and 237/248). Journal article conclusions were extracted from the article abstract. * Number 353 was changed to 391 since there appears to be a typographical error in the FDA medical review (the conclusion regarding trial 353 is included on another page of the review).*

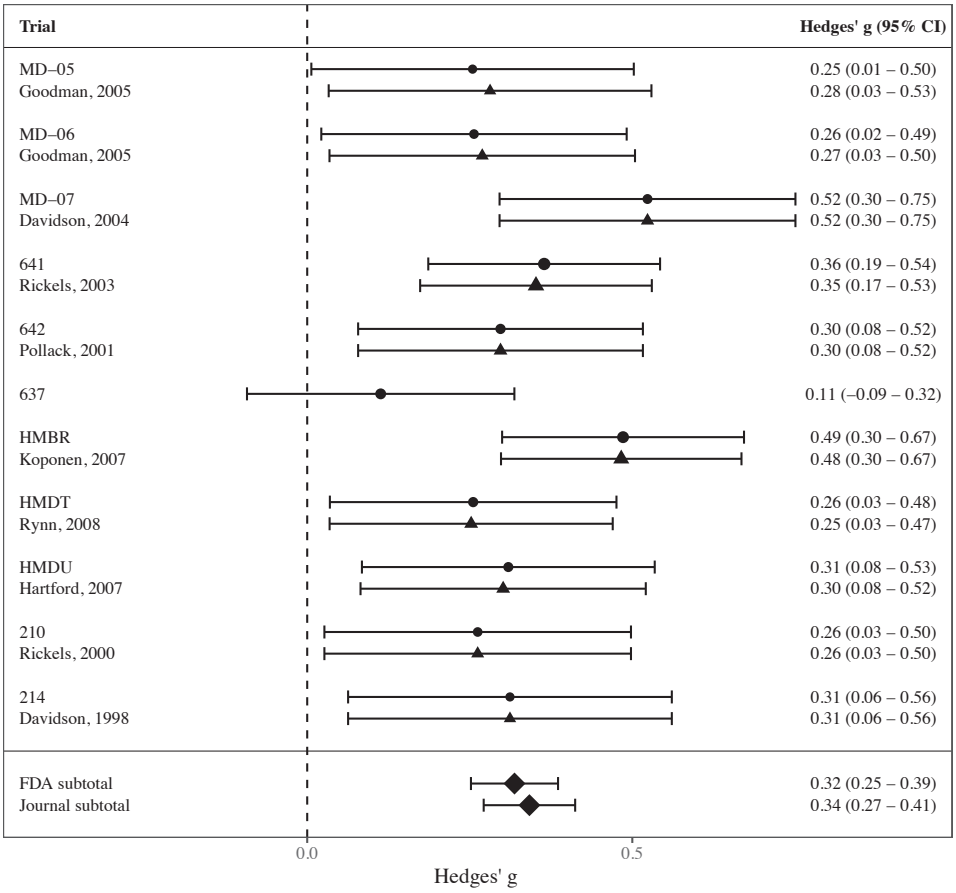

Figure 2.4: *Forest plot for GAD. Circles indicate FDA effect sizes, while triangles indicate the matching journal effect sizes (for published trials).*

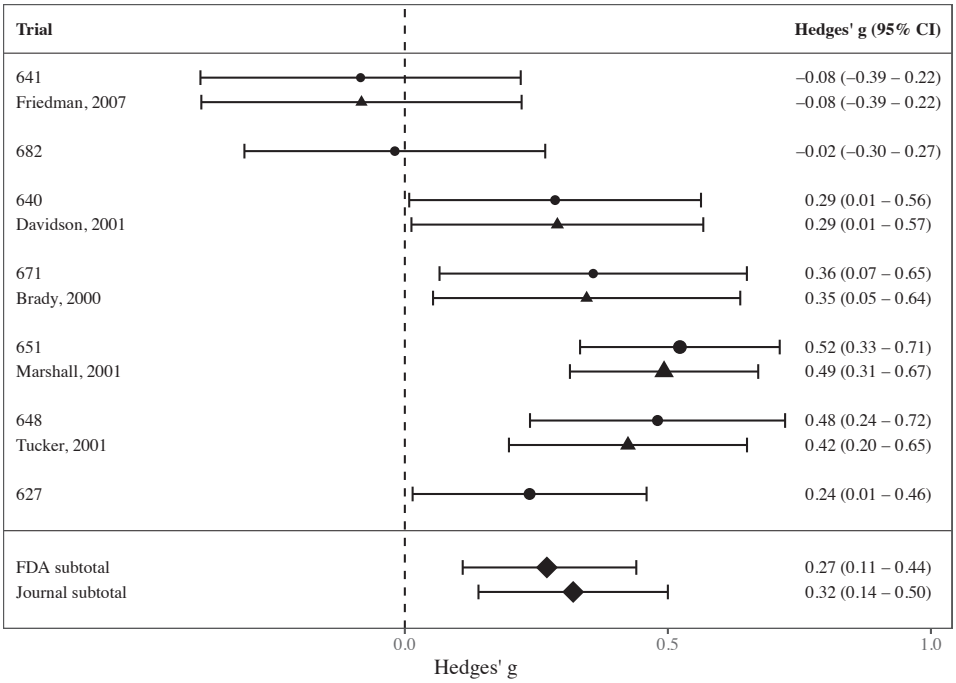

Figure 2.5: *Forest plot for PTSD. Circles indicate FDA effect sizes, while triangles indicate the matching journal effect sizes (for published trials).*

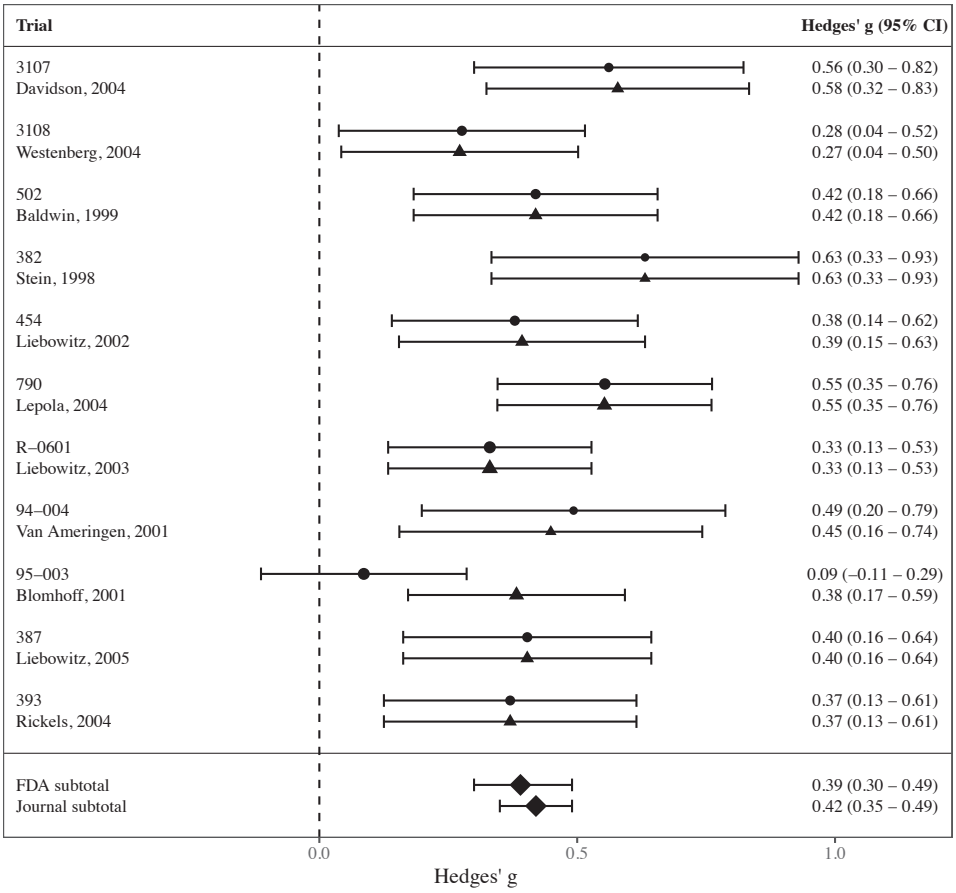

Figure 2.6: *Forest plot for SAD. Circles indicate FDA effect sizes, while triangles indicate the matching journal effect sizes (for published trials).*

Chapter 3

Bias in the reporting of harms in clinical trials of second-generation antidepressants for depression and anxiety

Ymkje Anna de Vries, Annelieke M. Roest, Lian Beijers,
Erick H. Turner, Peter de Jonge

European Neuropsychopharmacology (2016), 26, 1752 - 1759

Abstract

Background: Previous research has shown that reporting bias has inflated the apparent efficacy of antidepressants. We investigated whether apparent safety was also affected.

Methods: We included 133 trials, involving 31,296 patients, of second-generation antidepressants for the treatment of major depressive disorder (MDD) or anxiety disorders, obtained from Food and Drug Administration (FDA) reviews. We extracted data on overall discontinuation, discontinuation due to adverse events, and serious adverse events (SAEs). Meta-analysis was used to compare discontinuation rates between FDA reviews and matching journal articles, while SAEs were compared qualitatively.

Results: The odds ratio for overall discontinuation, comparing drug to placebo, was 1.0 for both sources, while that for discontinuation due to adverse events was 2.4 for both sources. Seventy-seven of 97 (79%) journal articles provided incomplete information on SAEs; sixty-one (63%) articles made no mention of SAEs at all. Of 21 articles which could be compared to the FDA, only 6 (29%) had full reporting without discrepancies. Nine (43%) articles reported a discrepant number of SAEs. Descriptions were absent or discrepant in 6 (29%) additional articles, even for important SAEs such as suicide attempts.

Conclusions: Reporting bias has not affected average discontinuation rates over trials. However, SAE reporting is not only very poor, with over half of articles failing to discuss SAEs altogether, but discrepancies between the FDA and articles were common and often led to a more favorable drug-placebo comparison. These findings suggest that journal articles are an unreliable source of data on SAEs in antidepressant trials.

Introduction

A significant fraction of all studies are never published in peer-reviewed journals [38]. Even within the subset of studies that are published, the (primary) analyses and outcomes reported in journal articles frequently deviate from the protocol [103, 186]. As a consequence, statistically significant (positive) studies or outcomes are more likely to be published than non-significant (negative) studies [38] or outcomes [103].

While it is often difficult to assess the presence of reporting bias, the United States Food and Drug Administration (FDA) maintains an independent results database for drug trials, which can be used to examine the presence of reporting bias within a set of trials [105]. This database has previously been used to assess reporting bias in trials of antipsychotics for schizophrenia [104] and antidepressants for major depressive disorder (MDD) [19] and anxiety disorders [20].

Second-generation antidepressants have been found to be effective for MDD [19] and anxiety disorders [113, 116, 119, 122, 123, 187]. They are considered to have a favorable risk-benefit profile and hence are widely prescribed [109]. While both studies examining the FDA database of antidepressant trials confirmed their efficacy for MDD and anxiety disorders, they also revealed substantial reporting bias [19, 20]. Although nearly all published trials (94 – 96%) reported positive results, only 51% of all submitted trials for MDD, and 72% of those for anxiety disorders, were judged to be positive by the FDA. As a consequence of reporting bias, the effect size of antidepressant treatment was overestimated by 32% and 15% for MDD and anxiety disorders, respectively.

An accurate assessment of the risk-benefit ratio of antidepressants requires an unbiased understanding of safety as well as efficacy, but this other side of the coin has not, thus far, been examined as comprehensively. Previous research has indicated that reporting of harms in journal articles is incomplete and inadequate in various medical fields [54, 55, 56], including psychiatry [53].

The case of reboxetine demonstrates the impact that reporting bias can have on apparent safety as well as efficacy: inclusion of unpublished data not only shifted the difference in efficacy between reboxetine and placebo from significant to non-significant, but it also showed that reboxetine was significantly inferior to placebo in terms of selected harm outcomes, while the published trials suggested they were equivalent [101]. Poor reporting of harms has also been found in trials of two other antidepressants (sertraline and duloxetine) and several antipsychotics, with serious adverse events (SAEs) not always reported fully or accurately in journal articles [57, 58].

The work on antidepressant trials was limited to relatively recent trials of three antidepressants, and only the reboxetine study quantified the possible impact of bias on an important harm outcome, discontinuation from the trial. In the present study, we assessed the presence of reporting bias, and its impact on several harm outcomes, within

a comprehensive set of trials of second-generation antidepressants for both MDD and anxiety disorders.

Methods

Data from FDA reviews and journal articles

We previously obtained FDA reviews of second-generation antidepressants approved for MDD [19] and/or anxiety disorders [20] (specifically generalized anxiety disorder (GAD), social anxiety disorder (SAD), obsessive-compulsive disorder (OCD), post-traumatic stress disorder (PTSD) and panic disorder (PD)). We defined second-generation antidepressants as including selective serotonin reuptake inhibitors (SSRIs), serotonin-norepinephrine reuptake inhibitors (SNRIs), as well as other antidepressants (specifically mirtazapine, bupropion, and nefazodone) approved between 1987 and 2008.

From these reviews, we identified all phase 2/3 short-term clinical trials registered with the FDA and conducted in pursuit of marketing approval. For MDD, we identified 74 trials of 12 drugs; for GAD, 11 trials of 4 drugs; for SAD, 11 trials of 5 drugs; for OCD, 13 trials of 5 drugs; for PTSD, 7 trials for 2 drugs; and for PD, 17 trials for 5 drugs. Two of the PD trials were not included in our previous analysis [20], as we did not receive the FDA review containing these trials in time. Hence, we included a total of 133 trials, consisting of data from 31,296 participants, of whom 18,904 were treated with antidepressants and 12,392 with placebo.

We conducted an extensive search of the published literature to identify journal articles corresponding to these FDA-registered trials, as described previously [19, 20]. A total of 97 publications were identified, covering 102 (77%) of 133 trials: 51 for MDD (including 1 publication covering 2 trials), 9 for GAD (1 publication covering 2 trials), 11 for SAD, 9 for OCD (1 publication covering 2 trials), 5 for PTSD, and 12 for PD (1 publication covering 3 trials).

For each trial, we extracted the following data from FDA reviews and corresponding journal articles, separately for each treatment group: sample size, number and proportion of patients discontinuing, number and proportion of patients discontinuing due to adverse events specifically, and the number and nature of serious adverse events (SAEs). SAEs are defined as any adverse event that results in death, hospitalization, disability or permanent damage, a birth defect, or any other life-threatening situation. Individual trial protocols may, however, define additional adverse events as serious adverse events. Both SAEs occurring during the administration of a drug and those occurring within a specified period (usually 30 days) after the last dose (post-therapy SAEs) must be reported to the FDA.

For trials which subdivided discontinuation due to adverse events into subcategories, we counted all participants who discontinued due to side effects, laboratory abnormalities, test findings, suicide attempts, suicide, and other causes of death, whether considered drug-related or not. When the journal article stated that no SAEs attributable to the drug were observed, we counted the article as reporting zero SAEs for the drug group.

We extracted data from (in order of preference) the safety population (all randomized patients who took at least one dose of medication or placebo), the randomized population, or the intention-to-treat efficacy population (all patients who took at least one dose of medication or placebo and who had at least one post-baseline efficacy evaluation). Extraction was performed independently by two investigators (YV and AR for depression and YV and LB for anxiety), with disagreements resolved by consensus. All remaining discrepancies between journal articles and FDA reviews were double-checked for possible errors.

Statistical analysis – discontinuation rates

For fixed-dose (multiple dose) studies, we calculated a combined sample size, and a combined number of patients discontinuing, for the various antidepressant arms. We then calculated the weighted mean discontinuation rate (overall [i.e., for any reason] and due to adverse events) for the drug and placebo group over all trials, as well as specifically per disorder.

We conducted four random-effects meta-analyses, as we had two data sources (FDA and journal articles) and two discontinuation rates (overall and due to adverse events). Restricted maximum likelihood (REML) was used as the estimation method. In case of empty cells (zero patients discontinuing in a group), 0.5 was added to the empty cell. Meta-regression was used to further investigate the impact of data source (FDA versus journal). We also conducted subgroup analyses for the different disorders. The metafor package (version 1.9-5) in R (version 3.1.3) was used for these analyses.

Analyses of serious adverse events

We first examined the availability of SAE data for the journal articles and the FDA reviews. For the journal articles that mentioned SAEs, and for which a comparison with the FDA review could be made, we determined whether there was a discrepancy in the number or description of SAEs in either the placebo or the drug group. In case of discrepancies, we further examined the nature of the discrepancy and whether it would lead to a more favorable drug-placebo comparison.

Results

Discontinuation rates

The average discontinuation rate (weighted by sample size) in the placebo group was 30.5% according to the FDA versus 30.1% according to the journal articles. For the antidepressant groups, the average discontinuation rate was 31.9% according to the FDA and 30.4% according to the journal articles (Table 3.1).

Random-effects meta-analysis of the FDA data yielded an odds ratio (OR) of discontinuation for the drug group compared with placebo of 1.02 (95% CI: 0.96 – 1.08), indicating no significant difference in odds of overall discontinuation between the drug and placebo groups. Similarly, random-effects meta-analysis of the journal data gave an OR of 0.98 (95% CI: 0.92 – 1.06). Lack of significant bias in journal articles compared to FDA was confirmed by meta-regression (p = 0.47) (Table 3.2).

For discontinuation due to adverse events, the average discontinuation rate was 5.2% for placebo according to the FDA and 5.0% according to the journal articles, while it was 12.6% for antidepressants according to the FDA and 12.3% according to journal articles. Random-effects meta-analysis of the FDA data yielded an odds ratio of 2.39 (95% CI: 2.11 – 2.70), indicating a significantly higher risk of discontinuation due to adverse events in the drug group. The journal articles gave similar results, with an OR of 2.42 (95% CI: 2.11 – 2.78), which was not significantly different from the FDA result (p = 0.94) (Table 3.2).

Table 3.1: *Discontinuation rates*

	Overall discontinuation (%)				Discontinuation due to AEs (%)			
	FDA		Journal		FDA		Journal	
	Placebo	Drug	Placebo	Drug	Placebo	Drug	Placebo	Drug
MDD	35.0	34.5	34.5	32.7	5.6	13.1	5.0	12.7
GAD	24.2	28.6	25.5	29.8	4.4	13.7	4.9	14.1
SAD	30.7	33.3	31.2	32.8	4.0	14.5	3.4	14.8
OCD	24.5	28.3	25.6	25.1	5.4	12.0	5.6	11.3
PTSD	32.1	32.5	31.3	34.3	7.4	11.5	6.9	12.0
PD	25.5	26.6	24.8	25.1	5.4	8.9	5.4	8.5
Total	30.5	31.9	30.1	30.4	5.2	12.6	5.0	12.3

Sample-size weighted average rates of overall discontinuation and discontinuation due to adverse events (AEs) for the placebo and drug groups according the Food and Drug Administration (FDA) and journal articles. GAD: generalized anxiety disorder; MDD: major depressive disorder; OCD: obsessive-compulsive disorder; PD: panic disorder; PTSD: post-traumatic stress disorder; SAD: social anxiety disorder.

Table 3.2: *Odds ratio of discontinuation*

	Overall discontinuation		Discontinuation due to AEs	
	FDA	Journal	FDA	Journal
MDD	0.92 (0.85-0.99)	0.87 (0.79-0.96)	2.29 (1.93-2.73)	2.42 (1.98-2.96)
GAD	1.25 (1.07-1.46)	1.26 (1.07-1.50)	3.06 (2.32-4.03)	2.85 (2.15-3.76)
SAD	1.14 (0.91-1.44)	1.08 (0.87-1.31)	3.75 (2.47-5.70)	4.33 (2.97-6.31)
OCD	1.24 (0.96-1.60)	0.99 (0.77-1.27)	2.28 (1.62-3.19)	2.23 (1.55-3.19)
PTSD	0.99 (0.82-1.21)	1.09 (0.87-1.36)	1.67 (1.11-2.51)	1.72 (1.18-2.52)
PD	1.06 (0.88-1.26)	0.99 (0.80-1.22)	1.57 (1.18-2.14)	1.51 (1.04-2.17)
Total	1.02 (0.96-1.08)	0.98 (0.92-1.06)	2.39 (2.11-2.70)	2.42 (2.11-2.78)

Odds ratio (95% confidence interval) of overall discontinuation and discontinuation due to adverse events (AEs) for the drug group compared to the placebo group according the FDA and journal articles. GAD: generalized anxiety disorder; MDD: major depressive disorder; OCD: obsessive-compulsive disorder; PD: panic disorder; PTSD: post-traumatic stress disorder; SAD: social anxiety disorder.

When we examined the disorders separately, no bias was apparent for any of the included disorders (all p-values >0.23). For GAD, overall discontinuation was significantly higher for the drug group than for the placebo group (OR = 1.25, 95% CI: 1.07 – 1.46, p = 0.004), while it was significantly lower for the drug group for MDD (OR = 0.92, 95% CI: 0.85 – 0.99, p = 0.026).

Discontinuation due to adverse events was significantly higher in the drug group for all disorders, with the magnitude of the difference ranging from an OR of 1.57 (for PD) to 3.75 (for SAD) on the basis of FDA data (Table 3.2).

Serious adverse events

Data on serious adverse events was frequently missing from both FDA reviews and journal articles. Out of 133 trials, SAE data was missing in the FDA review for 57 trials (43%). Nearly all of these were older trials for MDD, where data for 56 out of 74 (76%) trials was missing. For an additional 24 trials (18%), trial-level data on SAEs was not available in the FDA review, as the data was pooled over all pivotal trials included in the review.

Of 97 journal articles (covering 102 trials), only 36 (37%) mentioned SAEs at all: complete information was provided in 20 (21%) articles, while 16 (16%) articles had incomplete reporting (e.g. mentioning the number of SAEs without providing descriptions, or giving information for the drug group only).

For 15 of the 36 articles, there was insufficient information in the FDA review to perform a direct comparison. Of the remaining 21 articles, for which study-level SAE data was provided in the FDA reviews, only 6 (29%) articles had full reporting with no discrepancy.

Table 3.3: *Discrepancies in number of reported SAEs*

Disorder	Trial	FDA reporting	Journal reporting
MDD	84023 [188]	2 suicides in mirtazapine group vs. none in placebo group	"No clinically important SAEs ... attributable to mirtazapine were seen"
GAD	HMDT [135]	Severe anxiety and post-therapy death due to asphyxiation (choking)	Anxiety only
SAD	387 [158]	Event listed as "other event of clinical interest"	Event listed as SAE
	393 [159]	Event listed as "other event of clinical interest"	Event listed as SAE
	454 [153]	1 SAE in paroxetine group (brain edema due to car accident)	"No SAEs attributable to paroxetine treatment"
	94-004 [156]	Pregnancy with post-treatment abortion	Not mentioned
PD	399 [146]	1 SAE in placebo group and 5 SAEs in venlafaxine groups	8 SAEs in placebo group and 5 SAEs in venlafaxine groups
	391 [148]	4 SAEs in placebo group and 6 SAEs in drug group	Additional SAE of anxiety in placebo group
	495 [142]	Seizure	Not mentioned

FDA: Food and Drug Administration; GAD: generalized anxiety disorder; MDD: major depressive disorder; PD: panic disorder; SAE: serious adverse event.

In 9 out of 21 (43%) articles there were discrepancies in the reported number of SAEs, which led to a smaller or reversed drug-placebo difference and thus a more favorable drug-placebo comparison in 7 cases (Table 3.3).

Two articles reported additional SAEs in the drug group, which had been classified as "other clinical events of interest" by the FDA. Additional SAEs in the placebo group, which were not noted in the FDA review, were reported by two articles. Two other articles noted that "no SAEs attributable to [drug]" were seen, even though SAEs did occur in the drug group according to the FDA, including two suicides. Post-therapy SAEs were omitted in two other articles, while one article omitted an SAE of seizure in the drug group that had been noted by the FDA.

In 6 of 21 (29%) articles, the reported numbers of SAEs agreed, but a description of the SAEs was either (partly) missing in the journal article (5 articles) or differed from the FDA description of the SAE (1 article) (Table 3.4). The discrepancy in description involved an SAE that was described only as "emotional lability" in the journal article,

Table 3.4: *Missing or discrepant SAE descriptions*

Disorder	Trial	FDA reporting	Journal reporting
MDD	448, 449 [189]	5 SAEs in placebo group (4 somatic SAEs; accidental overdose) and 15 SAEs in drug groups (5 somatic SAEs; 3 cases of unintended pregnancy; abortion; post-therapy depression; 3 cases of emotional lability; depression and emotional lability; manic reaction)	No description
SAD	3108 [150]	1 SAE in placebo group (nasal septum disorder)	No description
	502 [151]	"Emotional lability/intentional OD of paracetamol and aspirin"	"Emotional lability"
OCD	3103 [169]	2 SAEs in placebo group (neoplasm, unintended pregnancy) and 5 in drug group (suicide attempt, accidental injury, asthma, hostility, depression)	No description
PTSD	651 [163]	9 SAEs in paroxetine group (headache; accidental overdose; benign neoplasm; uterine neoplasm; unintended pregnancy; 2 cases of emotional lability; manic reaction)	9 SAEs noted, but only those thought possibly related to treatment specified (headache, accidental overdose)
	671 [162]	1 SAE in placebo group (hives) and 1 SAE in drug group (post-treatment suicide attempt)	No description

FDA: Food and Drug administration; MDD: major depressive disorder; OCD: obsessive-compulsive disorder; OD: overdose. PTSD: post-traumatic stress disorder; SAD: social anxiety disorder; SAE: serious adverse event.

while it was described as "emotional lability/intentional OD [overdose] of paracetamol & aspirin" by the FDA.

With regard to missing descriptions, one article specified only those SAEs thought to be related to drug, while four articles did not provide descriptions for any SAEs. In four of these five articles with missing descriptions, the corresponding FDA reviews showed that there was a preponderance of psychiatric SAEs specifically in the drug groups, including suicide attempt, hostility, depression, emotional lability, and manic reaction, while there were almost no such SAEs in the placebo group.

Discussion

Principal findings

We found no evidence for bias in the reporting of overall discontinuation rates or discontinuation rates due to adverse events. However, discontinuation rates were high, averaging approximately 30% in both placebo and drug groups, especially given the short (6 – 12 weeks) duration of most of the included trials. Furthermore, participants receiving an antidepressant were 2.4 times more likely to discontinue the trial due to adverse events than participants receiving placebo.

Reporting of SAEs was very poor. In 79% of all journal articles, SAE data was incomplete or missing. Almost two-thirds failed to mention SAEs entirely, and an additional 16% of articles provided incomplete information regarding SAEs. For instance, some articles provided only the number of SAEs, without any description. Given the idiosyncratic nature of SAEs, such numbers have little meaning. Other articles provided the number of SAEs for the drug group only. However, lacking information about the base rate of SAEs in the placebo group, readers cannot ascertain whether the rate (or nature) of SAEs in the drug group should be cause for concern. In some of these cases, the FDA review showed that the number of SAEs in the placebo group was zero, but this was not always the case and hence cannot be assumed.

Furthermore, where a direct comparison with the FDA data was possible, discrepancies were frequent and usually led to a more favorable comparison between drug and placebo in the journal article. Discrepancies in the number of reported SAEs most commonly originated from the omission of post-therapy SAEs or the omission of events that were judged to be unrelated to treatment. However, such judgments are made subjectively by site investigators. As they are blind to treatment assignment, and the safety profile of a new drug cannot be known until the entire clinical trials program has been completed, it is impossible for investigators to determine causality.

As evidence that such judgments lead to undue omission of SAEs, in the current study, one article concluded that "no clinically important serious adverse events or side effects attributable to mirtazapine were seen" [188], even though, according to the FDA review, 2 (of 59) patients treated with mirtazapine died by suicide. Other articles neglected to provide a description for some or all SAEs, while the FDA data revealed a greater number of psychiatric adverse events specifically in the drug group. Such psychiatric adverse events are of particular concern, since antidepressants have been most controversially associated with suicidality (especially in children and young adults) [190, 191, 192] and aggression or violence [192, 193, 194, 195].

Because SAEs are infrequent, it is often difficult to reliably determine whether they are due to a drug on the basis of an entire clinical trials program, much less a single

antidepressant trial. Given the small numbers, an excess of SAEs in the drug group may be a chance finding, so many authors may choose to omit these data from journal articles to avoid alarming clinicians. Although meta-analysis of a large set of trials can help compensate for the small sample size of individual trials, poor reporting hampers meta-analysis of harm outcomes by independent authors. Outcome reporting bias is said to be a major threat to the validity of systematic reviews of harm outcomes [196]. Space limitations imposed by journals might also be a reason to omit SAEs entirely or SAE descriptions specifically, particularly when combined with a generally greater interest in efficacy results than in safety results on the part of editors and readers.

Comparison with previous literature

Several studies have found that, compared to ClinicalTrials.gov, SAE reporting in corresponding journal articles is incomplete [197, 198, 199]. Discrepancies are common, and journal articles generally report fewer SAEs. Similar to our results, two of these studies found that one reason for discrepancies was that journal articles reported only SAEs they judged to be drug-related [197, 198].

Regarding antidepressants specifically, Maund and colleagues examined the reporting of harms and benefits in 9 trials of duloxetine for MDD (eight of which were also included in the current study) [57]. Consistent with our results for duloxetine for MDD specifically, they found no discrepancies in the reporting of discontinuations because of adverse events or SAEs, although SAE reporting was very incomplete.

Hughes and colleagues, examining trials of several psychotropic medications (including duloxetine and sertraline) also found that nearly half of all SAEs reported in trial summaries were not reported in the associated journal articles, including over half of all cases of death and suicide [58]. Where SAEs were reported, discrepancies were common. Among the reasons for discrepancies were not reporting SAEs occurring during follow-up and only reporting "drug-related" SAEs, both of which we also found.

"Emotional lability" was one of the most frequently encountered SAE terms in this study. In clinical trials, narrative descriptions of adverse events are coded to a preferred term by use of a medical dictionary [200], but it is often unclear what this term actually means. A recent re-analysis of patient-level data from a trial of paroxetine for MDD in children and adolescents showed that this term was used to code for suicidal ideation, self-harm, or suicide attempts [33]. Consistent with this, we observed a case in which the journal article mentioned only emotional lability, while the FDA additionally reported intentional overdose. Hughes and colleagues also found cases in which, compared to trial summaries, journal articles provided markedly less informative SAE descriptions (e.g. "worsening of the illness" compared to "suicidal ideation") [58]. Hence, using vague descriptions may be an additional way in which SAE reporting can be biased.

Strengths and limitations

Among the strengths of our study is our use of the independent FDA database, which permitted us to identify a complete cohort of pre-marketing trials. As a consequence, and in contrast to previous research, we were able to include a wider range of antidepressants. We also examined important harm outcomes beyond SAEs, providing a more complete overview of reporting on harms.

Our study was limited by the information missing from FDA reviews, particularly with regard to SAEs, which hampered our ability to perform direct comparisons between journal articles and the FDA. This was especially true for older trials of antidepressants for MDD. Although SAEs were almost certainly examined during the drug approval process, this information was missing from the drug application packages provided to us. Therefore, although some information might still be found on clinical trial registries, data on SAEs is unavailable to the public for a significant fraction of all trials performed to obtain marketing approval for second-generation antidepressants, which are currently prescribed to approximately 10% of the US population [109].

Another limitation is that we did not examine common adverse events. Previous work has found that many common adverse events are not reported in journal articles, as these often only report those above a certain frequency threshold (e.g. >10% incidence in the drug group) [57].

Furthermore, we examined specifically whether the same information could be extracted from journal articles as from FDA reviews, without examining more subtle biases. Previous studies, for example, have shown that much less space is devoted to reporting on harms than is devoted to reporting on efficacy [53, 54]. In this study, we found that reporting of the actual discontinuation rates was unbiased, but many journal articles conclude, in their abstract, that the antidepressant was "safe", "well-tolerated", or both, even though antidepressant-treated participants were, on average, 2.4 times more likely to discontinue due to adverse events than placebo-treated patients. This might be considered a form of spin, particularly if the abstract makes no mention of the actual discontinuation rates or occurrence of adverse events.

Conclusions

While reporting of discontinuation rates showed no bias, reporting of SAEs was very poor, and inconsistencies between journal articles and FDA reviews were common. Previous research has shown that adoption of the CONSORT checklist leads to better reporting of the design, participant flow and results for efficacy outcomes in a trial [201]; compliance with the CONSORT extension for harms should similarly improve reporting of harm outcomes [202]. However, information on harms in previously completed trials can only become accessible if pharmaceutical companies or investigators choose to release it. Our

results thus show that an accurate assessment of the risk-benefit ratio of many widely prescribed antidepressants is hampered by poor and biased reporting of SAEs.

Chapter 4

Hiding negative antidepressant trials by pooling them: the pooled-trials publication bias

Ymkje Anna de Vries, Annelieke M. Roest,
Erick H. Turner, Peter de Jonge

Submitted

Abstract

Background: Previous studies on reporting bias generally examined whether trials were published in stand-alone publications. In this study we investigated whether pooled-trials publications constitute a specific form of reporting bias. We assessed whether negative trials were more likely to be exclusively published in pooled-trials publications than positive trials. In addition, we examined the research questions, individual trial results, and conclusions presented in these articles.

Methods: Data from a cohort of 74 randomized controlled trials of 12 antidepressants were extracted from an earlier publication and the corresponding Food and Drug Administration reviews. A systematic literature search was conducted to identify pooled-trials publications.

Results: We found 86 pooled-trials publications that reported results of 20 (83%) of 24 trials not published in stand-alone publications. Relative to positive trials, not-positive trials were 9.5 times more likely to be published exclusively in pooled-trials publications ($p<0.001$). Ten (12%) of 86 publications had as primary aim to present data on the trial's primary research question (drug efficacy compared to placebo). Only 3 publications, reporting on 3 (15%) trials, presented individual efficacy data for the primary research question. Additionally, only 3 (3%) of 86 pooled-trials publications had a negative conclusion.

Interpretation: Compared to positive trials, negative trials of antidepressants for depression were much more likely to be reported exclusively in pooled-trials publications. Pooled-trials publications flood the evidence base with often-redundant articles that, instead of addressing the original primary research question, present (positive) results on secondary questions. Therefore, pooled-trials publications inflate the apparent efficacy of antidepressants.

Introduction

The presence of reporting bias has been demonstrated in many medical fields [38, 182, 203]. An important form of reporting bias is study publication bias, which occurs when trials with positive results are more likely to be published than those with negative results [102].

In studies on reporting bias, trials that are published exclusively in pooled-trials publications, which pool data from two or more trials, are usually regarded as unpublished [19, 104] or incompletely published [204]. In contrast, some pharmaceutical companies have argued that these trials have actually been published and should be counted as such [205].

Although pooled-trials publications have been found to provide new information, they may be particularly susceptible to bias, for example because it is often unclear how trials were selected for inclusion [175]. In addition, pooled-trials publications often have a research question that differs from the original research question of the included trials. For example, they may focus on differential efficacy in various patient subgroups, leading to substantial redundancy and the suggestion that many of these articles may represent 'salami publications' [174].

Another potential problem with pooled-trials publications is that, in contrast to positive trials, negative trials may often be published exclusively in pooled-trials publications. A study examining trials for five antidepressants approved between 1989 and 1994 found that positive trials were usually reported in stand-alone publications, while negative trials were frequently "bundled" (often with positive trials) into pooled-trials publications [27]. Consequently, pooled-trials publications may actually further bias the published literature, rather than helping to provide transparent access to trial results.

A previous meta-analysis found that 31% of antidepressant trials for major depressive disorder (MDD) remained unpublished [19]. However, pooled-trials publications were excluded from this study. In the present study, we use the trials included in this meta-analysis to investigate whether the practice of pooling trials for publication constitutes a specific form of reporting bias.

Our first aim was to assess whether unpublished antidepressant trials were actually published in pooled-trials publications and to determine how frequently negative trials were published exclusively in pooled-trials publications compared to positive trials. Our second aim was to evaluate how often the research question of pooled-trials publications corresponded to the original primary research question of the included trials and how often these publications reported individual trial results for this primary outcome. Finally, our third aim was to assess how often pooled-trials publications reached positive conclusions about the trial drug.

Methods

Trials submitted to the FDA

Information on phase 2/3 clinical trial programs for 12 second-generation antidepressants (bupropion sustained release, citalopram, fluoxetine, paroxetine controlled release [CR], duloxetine, escitalopram, mirtazapine, nefazodone, paroxetine, sertraline, venlafaxine immediate release [IR], venlafaxine extended release [XR]) was extracted from an earlier publication by Turner and colleagues [19] and the Food and Drug Administration (FDA) reviews used in that study.

These programs included 74 randomized, double-blind, placebo-controlled trials investigating the short-term treatment of major depression. Because pharmaceutical companies must preregister trials they intend to conduct in support of an application of marketing approval with the FDA, FDA reviews can be used as a registry and results database [105].

Consistent with Turner and colleagues [19], who extracted the FDA's regulatory decision (i.e., whether the primary endpoint(s) were judged to be positive or not), 38 trials were considered positive and 36 trials were considered not-positive in the current study.

We retrieved the references of 50 journal articles that reported the results of trials registered with the FDA from Turner and colleagues [19]. One article presented the pooled results of two identically designed trials of paroxetine CR [189]. This article was regarded as a pooled-trials publication in the current study. In addition, we found a matching stand-alone publication [206] for one trial (UK-06) regarded as unpublished by Turner and colleagues [19].

Trials published in pooled-trials publications only

We assessed whether trials that were unpublished according to Turner and colleagues [19] (i.e. trials not published in stand-alone form) were published in pooled-trials publications. Pooled-trials publications were defined as publications in which the individual patient data of two or more trials were analyzed. This included publications described as individual patient data meta-analyses, but it did not include other meta-analyses (based on aggregate data).

A systematic literature search was conducted in PubMed, EMBASE and the Cochrane Central Register of Controlled Trials, restricted to articles in English, until February 10, 2015. The search strategy included the name of the drug, terms related to depression and "placebo". Terms were customized to the search strategies of each database; for example, when searching PubMed for relevant citalopram publications the search syntax was: citalopram [Title] AND depress* [Title/abstract] AND placebo.

After identifying the pooled-trials publications, matches for each trial were identified using the following parameters: drug name, active comparator (when applicable), dosage groups, sample sizes, trial duration, and names of investigators. We included only pooled-trials publications for which individual trials could be matched to FDA-registered trials.

From each publication, we extracted the primary research question and whether individual trial results were reported. Research questions were categorized as "primary efficacy", i.e. the research question of the pooled-trials publication was the same as the original trial's primary research question (efficacy of the drug compared to placebo), or "not primary efficacy". The second category consisted of publications on secondary efficacy outcomes (e.g. anxiety, sleep problems, efficacy compared to an active comparator), predictors of efficacy (e.g. efficacy in subgroups, baseline severity), and other efficacy or safety outcomes.

In addition, AR and YV classified each pooled-trials publication as positive, neutral, or negative, based on the abstract. Publications were judged to be positive when the abstract claimed that the antidepressant was more effective than placebo or an active comparator, equal in efficacy to an active comparator, safer or better tolerated than placebo or an active comparator, equal in safety/tolerability to placebo or an active comparator, or simply "safe" or "well-tolerated". Publications were judged to be neutral when the publication was primarily methodological in orientation, for example assessing differences between analytical approaches. Differences were resolved by consensus.

Statistical analysis

We examined whether not-positive trials were more likely to be published exclusively in a pooled-trials publication than positive trials. Because of small cell sizes, p values were obtained using Fisher's exact test. In addition, risk ratios and their 95% confidence intervals are reported. Analyses were performed using Stata software, version 13.1.

Results

Pooled-trials publications

Twenty-four of 74 FDA-registered antidepressant trials were not published in stand-alone publications. Of these, 20 (83.3%) were included in 86 pooled-trials publications reporting results on 10 antidepressants (see Table 4.1).

As shown in Figure 4.1, of the 36 trials judged not-positive by the FDA, 18 (50%) were exclusively published in pooled-trials publications, compared to 2 (5.3%) FDA positive trials. Consequently, all positive trials were published in some form (either stand-alone or

pooled), as were 89% of the not-positive trials. Compared to positive trials, not-positive trials were 9.5 times more likely to be published exclusively in pooled-trials publications (risk ratio: 9.50; 95% CI: 2.37-38.06; Fisher's exact p<0.001).

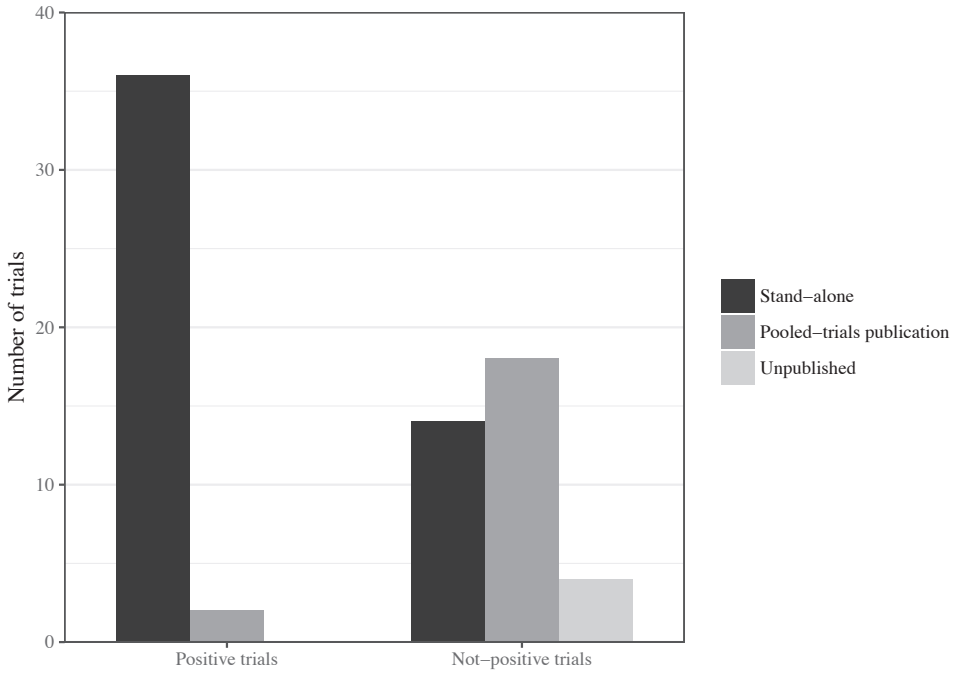

Figure 4.1: *Publication status of positive and not-positive FDA trials*

Research questions of trials published in pooled-trials publications and presentation of efficacy data

Ten out of 86 (11.6%) pooled-trials publications had the same research question as the included trial's original primary research question (drug efficacy compared to placebo) (Table 4.1). These 10 publications together included 7 (35%) of 20 trials published exclusively in pooled-trials publications. Only 3 (15%) of these publications presented individual efficacy data for the primary research question, reporting efficacy results for 3 not-positive trials.

Other pooled-trials publications reported on secondary efficacy outcomes (number of publications = 22; number of included trials = 9), predictors of efficacy (number of publications = 23; number of included trials = 8), other efficacy data (number of publications = 10; number of included trials = 5), or safety outcomes (number of publications = 21; number of included trials = 12) (Table 4.1).

Table 4.1: *Pooled-trials publications*

				Research question				
				Efficacy				
Drug	Trial	FDA	N	Primary	Secondary	Predictors	Other	Safety
Bupropion	205	NP	1	0 (0)	0	0	0	1
	212	NP	1	0 (0)	0	0	0	1
	Total		1	0 (0)	0	0	0	1
Citalopram	89306	NP	1	0 (0)	0	0	0	1
	Total		1	0 (0)	0	0	0	1
Duloxetine	HMAT-A	NP	38	4 (2)	6	12	4	12
	HMAQ-B	NP	31	2 (1)	6	12	1	10
	Total		38	4 (2)	6	12	4	12
Escitalopram	MD-02	NP	16	0 (0)	6	4	4	2
	Total		16	0 (0)	6	4	4	2
Mirtazapine	003-020	P	3	0 (0)	3	0	0	0
	003-021	NP	3	0 (0)	3	0	0	0
	003-003	NP	1	0 (0)	1	0	0	0
	003-008	NP	1	0 (0)	1	0	0	0
	Total		3	0 (0)	3	0	0	0
Nefazodone	7	NP	0	0 (0)	0	0	0	0
	004A	NP	0	0 (0)	0	0	0	0
	Total		0	0 (0)	0	0	0	0
Paroxetine	01-001	NP	2	0 (0)	0	0	0	2
(IR and CR)	03-003	NP	4	3 (0)	0	0	1	0
	07	NP	0	0 (0)	0	0	0	0
	09	NP	2	0 (0)	0	0	0	2
	UK-12	NP	2	0 (0)	0	0	0	2
	448	NP	2	1 (0)	0	1	0	0
	449	P	2	1 (0)	0	1	0	0
	Total		8	4 (0)	0	1	1	2
Sertraline	315	NP	1	0 (0)	0	1	0	0
	101	NP	1	0 (0)	0	0	0	1
	310	NP	0	0 (0)	0	0	0	0
	Total		2	0 (0)	0	1	0	1
Venlafaxine	303	NP	7	2 (1)	1	2	0	2
(IR and XR)	367	NP	12	1 (0)	6	3	1	1
	Total		17	2 (1)	7	5	1	2
All drugs	Total		86	10 (3)	22	23	10	21

The FDA column indicates the Food and Drug Administration decision (P: positive; NP: not-positive). N indicates the number of pooled-trials publications. For pooled-trials publications examining primary efficacy, the number in parentheses indicates the number of such publications that included individual trial results. CR: controlled release; IR: immediate release; XR: extended release.

Only three pooled-trials publications (3%) reported a negative conclusion (Table 4.2 in the Appendix). One of these publications examined the general safety profile of duloxetine and two examined the risk of suicidality with paroxetine treatment. All three concluded that the drug was associated with more adverse events than placebo. An additional 8 pooled trials publications (9%) had neutral conclusions (predictors of efficacy = 5; other efficacy = 3) while the remaining 75 (87%) were positively framed.

Discussion

To our knowledge, this study is the first to show that pooled-trials publication bias constitutes a specific form of reporting bias, which distorts the apparent efficacy of antidepressants. Although 24 of 74 antidepressant trials were not published in stand-alone articles, we showed that only four trials were completely unpublished, while the other 20 trials were included in pooled-trials publications. Trials lacking positive results were approximately 10 times more likely to be exclusively published in pooled-trials publications than trials with positive results.

Importantly, only 12% of all pooled-trials publications (including 7 (35%) of 20 trials) examined the original primary research question (efficacy of drug compared to placebo). Furthermore, individual trial results for this primary research question were provided in a pooled-trials publication for only 15% of trials. Finally, only 3% of pooled-trials publications had a negative conclusion. Therefore, although these trials have technically been published, the negative efficacy results are obscured, thus inflating the drugs' apparent efficacy.

For some drugs, particularly duloxetine, the number of pooled-trials publications was very high. Trials HMAT-A and HMAQ-B (which both had a negative result on the primary efficacy outcome) were included in 38 and 31 publications, respectively. Our study thus replicates a prior report on the 'salami slicing' of duloxetine trials [174].

Additionally, we found many pooled-trials publications for venlafaxine (17 publications for immediate- and extended-release combined) and escitalopram (16 publications). Several of these publications seemed redundant. For instance, three pooled-trials publications compared the efficacy of escitalopram to citalopram; three examined the effects of age and gender on the efficacy of venlafaxine; and eight compared the efficacy of venlafaxine to selective serotonin reuptake inhibitors.

It is noteworthy that duloxetine (approved in 2004) and escitalopram (approved in 2002) are the two newest antidepressants in our study, although venlafaxine ER was approved somewhat earlier in 1997. This suggests that the practice of pooling trials in many separate publications is a relatively new one, perhaps developing concurrently with physicians' growing skepticism of advertising and sales representatives and greater trust in peer-reviewed publications [207].

Consistent with this, a previous study examining five antidepressants approved between 1989 and 1994 identified at most six pooled-trials publications for a single drug [27]. In light of the growing concern that the medical literature may function as a marketing tool for pharmaceutical companies [207, 208, 209, 210], pooled-trials publications, consisting primarily of secondary analyses of previously collected data, may provide an easy and inexpensive way to keep a drug 'in the spotlight' and enhance its sales.

Others have also noted that antidepressant meta-analyses (including pooled-trials publications and aggregate data meta-analyses) are massively produced, frequently have some kind of industry involvement, and almost never include any negative statements if one or more authors are industry employees [211]. In our study, 76 (88%) of 86 pooled-trials publications had at least one author who was employed by a pharmaceutical company and, as noted, only three publications had negative conclusions. In all three cases, the publication concluded that the antidepressant was associated with more adverse events than placebo, a finding that is not unexpected.

A significant proportion of pooled-trials publications examined safety and tolerability. Because many adverse events occur infrequently, individual trials often lack sufficient statistical power for signal detection; pooling trials addresses this issue. The link between antidepressants and suicidality, for instance, was convincingly established only by pooling trials across indications and drugs [190, 191].

However, pooled-trials publications can also mislead. For instance, bupropion SR was only approved at dosages of 300 – 400 mg/day, but a pooled-trials publication assessing its safety and tolerability pooled all dosage groups (50 – 400 mg/day) [212]. Since adverse events are often dose-dependent, this publication is likely to paint an overly optimistic picture of the safety and tolerability of bupropion SR. Furthermore, there is ongoing concern that meta-analyses of harm outcomes may be particularly threatened by selective outcome reporting [196]. For non-systematic pooled analyses, this concern is further increased by the possibility of selective inclusion of trials [175].

Limitations

A limitation of the current study is that we may have missed some pooled-trials publications that actually did include FDA-registered trials, because some publications provided too little information to allow matching. However, this would not decrease the impact of pooled-trials publication bias, since individual trial results are never included in these publications.

A second limitation is that we did not count pooled-trials publications that did report individual trial results for the original primary outcome but focused on a secondary research question, as we felt that it was unlikely that these results would be found by researchers or clinicians interested in the primary outcome. However, such publications

were few in number, so the impact of this is minor.

A final limitation of this study is that we assessed the presence of pooled-trials publication bias in a narrow field, namely antidepressant trials for MDD. Nevertheless, it would not be surprising to find such bias in other fields of medicine, because reporting bias has been found to occur throughout psychiatry [20, 33, 176], medicine [38, 182, 203], and science in general [34].

Conclusions

Although meta-analyses on reporting bias have been criticized for their decision to exclude pooled-trials publications, our study shows that these publications are biased toward positive conclusions. As these publications rarely include individual trial results, they appear to serve primarily to heighten the (positive) visibility of a drug, rather than to clearly and transparently report negative trial results.

Journal editors could request that pooled-trials publications also present individual trial results and reference articles that present the primary efficacy results for all included trials. Additionally, editors, peer reviewers, and readers should be aware of the potential for bias and redundancy with pooled-trials publications [174] and perhaps ask whether they enhance or merely distort and bloat the evidence base.

In summary, the practice of pooling trials increases the apparent efficacy of antidepressants by flooding the literature with publications that highlight positive results and obscure negative results. Together with study publication bias, selective outcome reporting, and spin, pooled-trials publication bias is a form of reporting bias that should be taken into account in future research.

Appendix

Table 4.2: *Supplemental table of studies*

Publication	Included trials	Research question	Abstract
Bupropion			
Settle (1999) [212]	205, 212	Safety	Positive
Citalopram			
Pedersen (2006) [213]	89306	Safety	Positive
Duloxetine			
Bailey (2006) [214]	HMAT-A, HMAQ-B	Predictors of efficacy	Positive
Bech (2006) [215]	HMAT-A	Other efficacy	Positive
Brecht (2008) [216]	HMAT-A, HMAQ-B	Secondary efficacy	Positive
Brunton (2010) [217]	HMAT-A, HMAQ-B	Safety	Negative
Cookson (2006) [218]	HMAT-A, HMAQ-B	Secondary efficacy	Positive
Delgado (2005) [219]	HMAT-A	Safety	Positive
Dodd (2014) [220]	HMAT-A, HMAQ-B	Predictors of efficacy	Neutral
Dunner (2003) [221]	HMAT-A, HMAQ-B	Secondary efficacy	Positive
Dunner (2005) [222]	HMAT-A, HMAQ-B	Safety	Positive
Fishbain (2008) [223]	HMAT-A	Other efficacy	Positive
Greist (2004) [224]	HMAT-A, HMAQ-B	Safety	Positive
Gueorguieva (2011) [225]	HMAT-A, HMAQ-B	Primary efficacy	Positive
Harada (2015) [226]	HMAT-A	Primary efficacy	Positive
Hudson (2005) [227]	HMAT-A, HMAQ-B	Safety	Positive
Kornstein (2006) [228]	HMAT-A, HMAQ-B	Predictors of efficacy	Positive
Lewis-Fernandez (2006) [229]	HMAT-A, HMAQ-B	Predictors of efficacy	Positive
Mallinckrodt (2003) [230]	HMAT-A, HMAQ-B	Secondary efficacy	Positive
Mallinckrodt (2004) [231]	HMAT-A, HMAQ-B	Other efficacy	Neutral
Mallinckrodt (2005) [232]	HMAT-A, HMAQ-B	Predictors of efficacy	Positive
Mallinckrodt (2006) [233]	HMAT-A	Primary efficacy	Positive
Mallinckrodt (2008) [234]	HMAT-A, HMAQ-B	Secondary efficacy	Positive
Nelson (2005) [235]	HMAT-A, HMAQ-B	Predictors of efficacy	Positive
Nelson (2006) [236]	HMAT-A	Safety	Positive
Nelson (2010) [237]	HMAT-A, HMAQ-B	Predictors of efficacy	Positive
Nelson (2011) [238]	HMAT-A, HMAQ-B	Predictors of efficacy	Neutral
Nelson (2013) [239]	HMAT-A, HMAQ-B	Predictors of efficacy	Positive
Nemeroff (2002) [240]	HMAT-A, HMAQ-B	Primary efficacy	Positive
Perahia (2005) [241]	HMAT-A, HMAQ-B	Safety	Positive
Perahia (2006) [242]	HMAT-A, HMAQ-B	Predictors of efficacy	Positive
Pritchett (2007) [243]	HMAT-A	Other efficacy	Positive
Schacht (2014) [244]	HMAT-A, HMAQ-B	Predictors of efficacy	Neutral
Shelton (2007) [245]	HMAT-A, HMAQ-B	Predictors of efficacy	Positive
Stewart (2006) [246]	HMAT-A, HMAQ-B	Safety	Positive
Thase (2005) [247]	HMAT-A, HMAQ-B	Safety	Positive
Thase (2007) [248]	HMAT-A, HMAQ-B	Secondary efficacy	Positive

continued

Table 4.2: *Supplemental table of studies*

Publication	Included trials	Research question	Abstract
Viktrup (2004) [249]	HMAT-A, HMAQ-B	Safety	Positive
Wernicke (2007) [250]	HMAT-A, HMAQ-B	Safety	Positive
Wise (2006) [251]	HMAT-A, HMAQ-B	Safety	Positive
Escitalopram			
Baldwin (2007) [252]	SCT-MD-02	Safety	Positive
Baldwin (2009) [94]	SCT-MD-02	Other efficacy	Neutral
Bandelow (2007) [253]	SCT-MD-02	Secondary efficacy	Positive
Demyttenaere (2008) [254]	SCT-MD-02	Secondary efficacy	Positive
Demyttenaere (2011) [255]	SCT-MD-02	Other efficacy	Neutral
Gorman (2002) [256]	SCT-MD-02	Secondary efficacy	Positive
Kennedy (2009) [257]	SCT-MD-02	Secondary efficacy	Positive
Kilts (2009) [258]	SCT-MD-02	Predictors of efficacy	Positive
Lader (2005) [259]	SCT-MD-02	Secondary efficacy	Positive
Lam (2006) [260]	SCT-MD-02	Predictors of efficacy	Positive
Llorca (2005) [261]	SCT-MD-02	Other efficacy	Positive
Papakostas (2011) [262]	SCT-MD-02	Predictors of efficacy	Positive
Pedersen (2005) [263]	SCT-MD-02	Safety	Positive
Stein (2006) [264]	SCT-MD-02	Other efficacy	Neutral
Stein (2011) [265]	SCT-MD-02	Secondary efficacy	Positive
Wade (2006) [266]	SCT-MD-02	Predictors of efficacy	Neutral
Mirtazapine			
Bech (2001) [267]	003-020, 003-021	Secondary efficacy	Positive
Fawcett (1998) [268]	003-003, 003-008, 003-020, 003-021	Secondary efficacy	Positive
Stahl (1997) [269]	003-020, 003-021	Secondary efficacy	Positive
Paroxetine IR and CR			
Carpenter (2011) [270]	448, 449, UK-12, 01, 09	Safety	Negative
Dunbar (1991) [271]	03-003	Primary efficacy	Positive
Feighner (1992) [272]	03-003	Primary efficacy	Positive
Feighner (1993) [273]	03-003	Primary efficacy	Positive
Kraus (2010) [274]	448, 449, UK-12, 01, 09	Safety	Negative
Montgomery (1992) [275]	03-003	Other efficacy	Positive
Dunner (2005) [276]	448, 449	Predictors of efficacy	Positive
Golden (2002) [189]	448, 449	Primary efficacy	Positive
Sertraline			
Berti (1995) [277]	315	Predictors of efficacy	Positive
Fisch (1992) [278]	101	Safety	Positive
Venlafaxine IR and XR			
Danjou (1995) [279]	303	Safety	Positive
Entsuah (1995a) [280]	303	Predictors of efficacy	Positive

continued

Table 4.2: *Supplemental table of studies*

Publication	Included trials	Research question	Abstract
Entsuah (1995b) [281]	303	Predictors of efficacy	Positive
Entsuah (2001) [282]	367	Predictors of efficacy	Positive
Entsuah (2002) [283]	367	Other efficacy	Positive
Gibbons (2012a) [76]	303, 367	Primary efficacy	Positive
Gibbons (2012b) [284]	303, 367	Safety	Positive
Mallick (2003) [285]	367	Secondary efficacy	Positive
Mendlewicz (1995) [286]	303	Primary efficacy	Positive
Nemeroff (2008) [287]	367	Secondary efficacy	Positive
Rudolph, (1998) [288]	303	Secondary efficacy	Positive
Shelton (2005) [289]	367	Secondary efficacy	Positive
Silverstone (2002) [290]	367	Predictors of efficacy	Positive
Stahl (2002) [291]	367	Secondary efficacy	Positive
Thase (2001) [292]	367	Secondary efficacy	Positive
Thase (2005) [293]	367	Predictors of efficacy	Positive
Trivedi (2004) [294]	367	Secondary efficacy	Positive

Chapter 5

Citation distortions in the literature on the serotonin-transporter-linked polymorphism and amygdala activation

Jojanneke A. Bastiaansen, Ymkje Anna de Vries, Marcus R. Munafò

Biological Psychiatry (2015), 78 (8), E35 - 36

Abstract

Selective citation may contribute to the persistence of strong beliefs in research findings for which the strength of evidence is questionable or declining. For the literature on the putative association between the serotonin transporter polymorphism (5-HTTLPR) and amygdala activation, differences between citation rates for positive and negative studies are obscured by large differences within the negative group: whereas studies that claim to have found an effect (in spite of negative findings indicated by a standardized meta-analytic approach) are cited as much as positive studies, studies that neither report nor claim the existence of an effect (i.e., "refutation" studies) are typically overlooked. Moreover, the majority of recent studies ignore the methodological issues raised by a meta-analysis and only mention the presence of a significant meta-analytic effect. Researchers should focus on the nuance and caveats of previous work in their articles and be encouraged to publish and cite refutation studies.

Short report

A seminal finding in imaging genetics is that carriers of the short (S) allele of the serotonin-transporter-linked polymorphic region (5-HTTLPR) exhibit an increased amygdala response to negative emotional stimuli [295]. The original article by Hariri and colleagues has been cited over 1,000 times since its publication in 2002.

Although meta-analyses have shown a statistically significant (but small) effect across published studies, the validity of these findings is undermined by the presence of publication bias [296, 297]. Moreover, the strength of evidence has declined over time [298, 297]. However, the strength of belief does not seem to have decreased comparably. For instance, a recent review maintained that up to 5% of differences in amygdala activation can be explained by variation in the 5-HTTLPR [299].

One factor that may contribute to the persistence of belief in an effect is preferential citation of positive studies [44]. For the network of studies reported in the most recent meta-analysis on 5-HTTLPR and amygdala activation [297], citation differences between positive (n = 10) and negative studies (n = 15), although present, are not very pronounced. While 40% of studies are positive, they receive 55% of within-network citations and 67% of citations via Web of Science (49% excluding Hariri and colleagues, 2002 [295]). A positive study is cited, on average, by 39% (SD = 32%) of subsequent studies in the network, and negative studies are cited by 25% (SD = 24%). In Web of Science, average yearly citation rates for negative and positive studies are 11 (SD = 11) and 24 (SD = 32) times, respectively, with the latter declining to 15 (SD = 17) times when Hariri and colleagues (2002) [295] is excluded.

However, citation rates of negative studies can be confounded by studies with inflated claims or "spin" in their abstracts. Spin is the (intentional or unintentional) use of reporting strategies to emphasize the presence of an effect, for instance by focusing on statistically significant findings from subgroup analyses or secondary outcomes [178].

It was previously shown that many of the studies in Murphy et al. (2013) [297] make stronger claims in their abstracts than is warranted by the reported data when a standardized analytical approach is employed [300]. Figures 5.1 and 5.2 illustrate that "claim" studies – that is, negative studies that claim to have found an effect, but for which a standardized analysis does not indicate statistically significance evidence - are cited comparably to positive studies.

In contrast, studies that neither report nor claim the existence of an effect (i.e., "refutation" studies) are overlooked. Refutation studies, for instance, are cited by only 14% of subsequent studies within the network (figure 5.2, left panel) and they receive only 4% of citations in Web of Science (Figure 5.1, right panel). Refutation studies appear to face a double difficulty in contributing to and changing the common perspective: not only is it hard to publish them, but they are also cited infrequently once published. Studies are

rewarded for making positive claims by higher citation rates, resulting in a literature that presents a distorted impression of the strength of evidence.

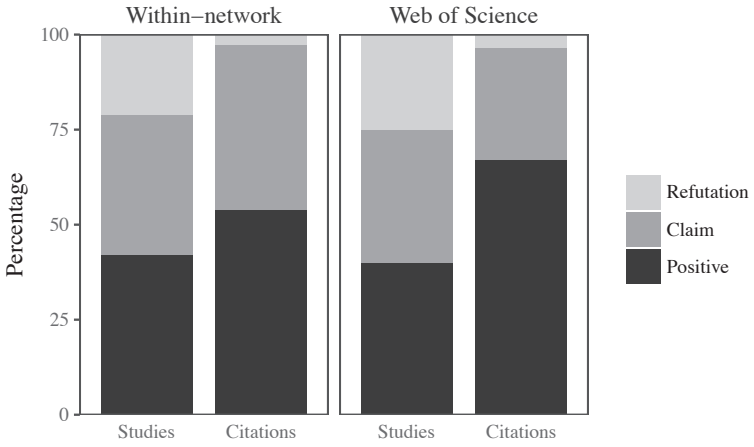

Figure 5.1: *Percentage of studies of each type (positive, claim, and refutation) and the percentage of citations received by each type of study for within-network citations and Web of Science citations. For within-network citations, the final study within the network was not included in the number of studies, as it could not have been cited within the network.*

Effect estimates by meta-analyses are not affected by spin and often swiftly become the new standard in the field. Although meta-analyses can potentially override the effects of citation distortion, they can also lead to further distortion when important issues they raise are neglected. Two independent raters coded whether the 37 peer-reviewed English-language articles citing Murphy et al. [297] (source: Google Scholar, November 2014) referred to these authors' concerns about issues of statistical power and publication bias, or only mentioned the presence of a statistically significant effect.

Three methodological articles did not address the outcome of the meta-analysis, and one did not provide enough information for coding. Of the remaining 33 articles, only seven reported Murphy et al.'s concerns, and one article discussed similar issues more broadly. In other words, 76% of recent studies cite the meta-analysis as evidence for the association without expressing concern regarding the validity of this conclusion.

Who and what is cited colors the common perception of an evidence base. For the association of 5-HTTLPR genotype with amygdala activation, we have shown that refutation studies are typically ignored and methodological concerns reported by a meta-analysis are often overlooked. Researchers should focus on the nuance and caveats associated with any result (including those derived from a meta-analysis) in their articles, and should be encouraged to publish and cite refutation studies. A recent study published individual (null) results together with an updated meta-analysis, an approach that might help increase the visibility of refutation studies [298]. Citation analysis of other topics could

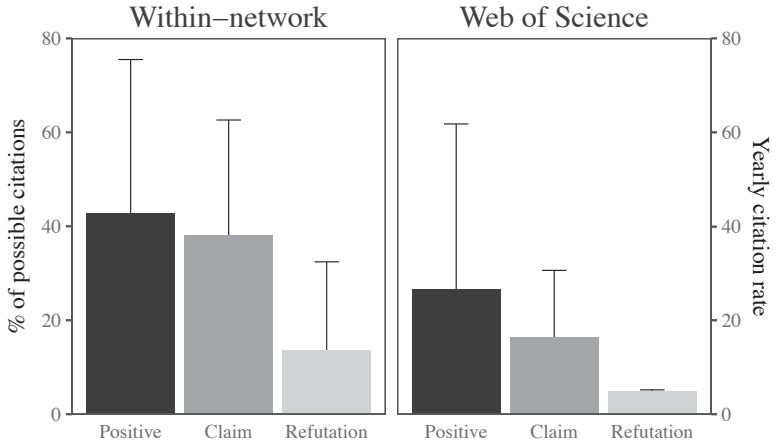

Figure 5.2: *% of possible within-network citations (±1 SD) and yearly Web of Science citation rate (±1 SD) for positive, claim, and refutation studies.*

help clarify why certain beliefs remain deeply rooted in the field and support researchers in distinguishing fad from fact.

Chapter 6

Citation bias and selective focus on positive findings in the literature on the serotonin transporter gene, life stress and depression

Ymkje Anna de Vries, Annelieke M. Roest, Minita Franzen,
Marcus R. Munafò, Jojanneke A. Bastiaansen

Psychological Medicine (2016), 46, 2971 - 2979

Abstract

Background: Caspi *et al.*'s 2003 report that 5-HTTLPR genotype moderates the influence of life stress on depression has been highly influential but remains contentious. We examined whether the evidence base for the 5-HTTLPR-stress interaction has been distorted by citation bias and a selective focus on positive findings.

Methods: Seventy-three primary studies were coded for study outcomes and focus on positive findings in the abstract. Citation rates were compared between studies with positive and negative results, both within this network of primary studies and in Web of Science. In addition, the impact of focus on citation rates was examined.

Results: Twenty-four (33%) studies were coded as positive, but these received 48% of within-network and 68% of Web of Science citations. The 38 (52%) negative studies received 42% and 23% of citations, respectively, while the 11 (15%) unclear studies received 10% and 9%. Of the negative studies, the 16 studies without a positive focus (42%) received 47% of within-network citations and 32% of Web of Science citations, while the 13 (34%) studies with a positive focus received 39% and 51% respectively, and the 9 (24%) studies with a partially positive focus received 14% and 17%.

Conclusions: Negative studies received fewer citations than positive studies. Furthermore, over half of the negative studies had a (partially) positive focus, and Web of Science citation rates were higher for these studies. Thus, discussion of the 5-HTTLPR-stress interaction is more positive than warranted. This study exemplifies how evidence-base-distorting mechanisms undermine the authenticity of research findings.

Introduction

Major depressive disorder (MDD) is a complex illness, caused by a combination of genetic and environmental risk factors [301]. One of the most robust risk factors for MDD is the experience of a stressor, such as a stressful life event or childhood abuse [302]. However, many who experience such a stressor do not develop depression. This individual variability has been suggested to be due, at least in part, to genetic variation [303].

In 2003, Caspi and colleagues reported that a polymorphism in the serotonin transporter gene (5-HTTLPR) moderates the relationship between life stress and depression: while carriers of at least one short (S) allele had a similar risk of depression as people homozygous for the long (L) allele in the absence of stress, S carriers were up to twice as likely to develop depression after stressful life events or childhood abuse [304].

This study has since been highly cited (>3,800 times, Web of Science, October 2015) and has become the seminal finding within the burgeoning field of gene-environment interactions (G×E). However, this finding also remains highly contentious. Even meta-analyses on this topic contradict each other, with some finding evidence of an effect [305, 306], while others do not [307, 308].

Many issues complicate the interpretation of G×E findings and replications, such as publication bias [309] and analytical flexibility [64, 310], which increases the chance of false-positives due to the multitude of analyses performed [311]. The likelihood of false-positives is further increased by low power and by the low prior probability of associations in candidate gene studies [309]. Although replication has been suggested as the solution to false-positive findings [312], many G×E studies are imprecise replications of the original finding, and a loose definition of replication may still permit propagation of false-positives [63].

Additionally, researchers may emphasize positive findings while downplaying negative findings. Within the randomized controlled trial literature, such reporting strategies, whether intentional or unintentional, that focus on positive (secondary) findings (in spite of non-significant results for the primary outcome) and that may distort the interpretation of results, are defined as 'spin' [20, 40]. A focus on positive findings has also been demonstrated in observational studies [300]. As a consequence, the published literature on a topic may appear more convincing than is justified by the strength of the evidence.

Selective citation may also affect the quality of the evidence base [44]. Statistically significant (positive) studies are cited more frequently than non-significant (negative) studies [42, 45, 313, 314], which may render non-supportive studies relatively invisible.

Citation bias and focus on positive findings can also work synergistically to hide negative results from view. A previous examination of citation patterns on a related topic, that of 5-HTTLPR and amygdala activation [315], showed that negative studies that had been spun were cited at a similar rate as positive studies, while negative studies that had not

been spun received almost no citations. The resulting invisibility of negative findings may create the impression that this effect has been proven beyond doubt, although meta-analyses have questioned its robustness [297, 298].

In the current study, we aimed to determine whether citation bias and selective focus on positive findings are also present in the literature on 5-HTTLPR, life stress, and depression. Achieving a better understanding of the etiology of depression is of vital importance to psychiatry, given the high burden of depression [316]. Distortion of the evidence base could mislead researchers and clinicians and thus pose a major obstacle to this goal.

Methods

Study selection

To establish the network of primary studies, we searched PubMed for the most recent meta-analysis on 5-HTTLPR, stress, and depression [305], which included 81 studies. For each study, we determined the outcome for the effect of interest (i.e., 5-HTTLPR x stress).

We included studies with continuous outcomes (e.g., score on a depression questionnaire), as well as studies with binary outcomes (e.g., depression diagnosis). We excluded studies in which the outcome was clearly a different construct than depression (e.g., cognitive dysfunction). Studies were included regardless of whether the 5-HTTLPR x stress interaction effect on depression was the primary outcome. No exclusion criteria were applied for stressors, which were very diverse.

Coding study outcomes

Coding was done in duplicate by two independent raters (YV and MF), and disagreements were resolved by discussion with AR and JB. Study outcome was coded as positive, negative, or unclear. We coded a study outcome as unclear if we could not determine whether the 5-HTTLPR x stress interaction was significant, for instance because only the p-value associated with a three-way interaction (e.g., 5-HTTLPR x stress x gender) was presented. Study outcome was coded as positive if the extracted p-value was <0.05, provided that the interaction was in the expected direction (i.e., S allele associated with increased depression), and as negative otherwise.

P-values were extracted according to a hierarchical decision tree. We first determined whether the design of the study was "exposed-only". In these studies, the entire sample was exposed to a stressor, such as a somatic illness. The effect of interest, in this case,

is not an interaction but the main effect of 5-HTTLPR. Hence, we extracted the p-value associated with the main effect for these studies. For all other studies, we determined whether a p-value was reported for a two-way interaction between 5-HTTLPR and stress, consistent with Caspi and colleagues (2003) [304].

If multiple relevant, independent outcomes or stressors were included in a study, we extracted all p-values. Following Sharpley and colleagues (2014) [305], we averaged these p-values to arrive at a conclusion. If multiple non-independent outcomes were given (e.g., a continuous symptom scale and a dichotomized version thereof), we only included the continuous outcome.

When studies provided p-values for both biallelic and triallelic genotyping, we erred towards coding a study as positive by selecting the smallest p-value, as it is unclear which genotyping approach should be preferred [317, 318, 319]. If both unadjusted and adjusted analyses were given, we also used the smallest p-value. We preferentially extracted the p-value of an overall test of interaction; however, if only post-hoc comparisons were available, we extracted the p-value associated with the SS vs LL homozygotes comparison.

Coding study abstracts

Two independent raters (YV and MF) coded the abstract of each study, and discrepancies were resolved by discussion with AR and JB. Abstracts were preferentially coded based upon their conclusions, but if these did not provide a clear statement, we used the results section of the abstract. In coding abstracts, we were interested in the way abstracts reported on how their findings reflected on the original result by Caspi and colleagues (2003) [304].

Abstracts were coded as positive if a claim was made that the results supported the existence and/or importance of the 5-HTTLPR x stress interaction. Abstracts were coded as partially supportive if a positive claim was made that was not directly related to the 5-HTTLPR x stress interaction (e.g., positive findings for a three-way interaction) or if the abstract mentioned findings for multiple outcomes or stressors and not all were positive. Abstracts that did not make a positive claim or that made an explicitly negative claim were coded as negative. If the abstract did not report on the effect of interest, the study was excluded (2 studies).

Citation outcomes

We examined citations both within the network of primary studies and outside of the network in the broader literature [315]. To examine within-network citations, we constructed a citation grid and marked for each study by which of the other included studies it was cited. Total citation counts for each study were calculated from the grid. To examine

out-of-network citations, we looked up the citation counts for each study on Web of Science (Core Collection, October 2015). To create non-overlapping outcomes, we pruned the within-network citations from the Web of Science citations. While within-network citations represent citations by other experts working within the 5-HTTLPR x stress field, Web of Science citations also include citations by researchers not directly involved in this area.

Analyses

For our citation analysis, we first compared the citations received by studies with positive, negative or unclear outcomes (irrespective of abstract coding). The sum of citations was calculated and the percentage of all citations received by studies with a given outcome was determined. In examining within-network citations, we excluded the most recent study, as it could not have been cited within the network. We also examined the study by Caspi and colleagues (2003) [304] separately, as we expected it to receive many citations.

To determine whether a (selective) focus on positive findings was present, we examined the number of negative studies with a negative abstract (studies without a positive focus), a partially supportive abstract (studies with a partially positive focus), or a positive abstract (studies with a positive focus). We then examined the impact of focus on citation rates by calculating the percentage of all citations to negative studies received by each type of negative study.

Within the network, we also examined whether positive studies, negative studies without positive focus, and negative studies with a (partially) positive focus showed different citation patterns, that is, whether positive studies were more likely to cite other positive studies and negative studies more likely to cite other negative studies.

We performed several sensitivity analyses. First, since older studies have had more opportunities to be cited, we re-examined citation rates based on measures taking into account publication year. For within-network citations, the percentage of subsequent studies citing a given study was calculated; for out-of-network citations, the yearly citation rate was calculated. Second, as the distribution of citations is right-skewed, we examined the median number of citations to each study type. Third, we recoded the outcome for studies with multiple relevant p-values based upon the smallest p-value. As it is often unclear what should be considered the primary outcome, we used average p-values in our main analysis; however, in some cases the smallest p-value may have been associated with the outcome considered most important by the authors, which is why we performed this sensitivity analysis.

Since the included studies form the total population of studies on the effect of interest, we used descriptive analyses rather than statistical tests [315], which are designed to generalize from a sample to a hypothetical larger population.

Results

Coding of studies and abstracts

We excluded ten of the 81 studies in Sharpley and colleagues (2014) [305]: eight studies were excluded because the outcome was not depression-related, no stressor was included, or the entire sample was depressed; one study was excluded because the abstract did not report on 5-HTTLPR; and one study was excluded because the abstract did not report on the depression outcome. Furthermore, we included two additional studies that had been excluded from the meta-analysis because the sample was a subset of those included in a later study.

Consequently, we included 73 studies, of which 24 studies were coded as positive, 38 studies as negative, and 11 studies as unclear in terms of outcome. Of the 11 unclear studies, four studies were coded as unclear because of the inclusion of three-way interactions in the model (e.g., with social support), while another study was coded as unclear because the 5-HTTLPR x stress interaction was only tested in males and females separately. Four studies were coded as unclear because the 5-HTTLPR x stress interaction was not tested (e.g., only the main effect of 5-HTTLPR in the different stress groups was tested). Finally, two studies were coded as unclear because we could not determine whether the (averaged) p-value was <0.05, as one p-value was given as "non-significant" while another was <0.05.

Inter-rater agreement was moderate (κ=0.49). Our agreement with Sharpley and colleagues (2014) [305] was good: within the subset of studies included in both Sharpley and colleagues (2014) and our own paper and that we coded as positive or negative (rather than unclear), the percentage of positive studies was 38% (23 out of 60) by both our coding and Sharpley's coding; coding was identical for 54 out of 60 (90%) papers.

Of the 73 studies, we coded 40 abstracts as positive, 16 abstracts as negative, and 17 abstracts as partially supportive. Inter-rater agreement for abstract coding was good (κ=0.71). A full table of studies with characteristics and coding is given in Table 6.1 in the Appendix.

Citations by study outcome

Figure 6.1 shows the percentage of citations to positive, negative, and unclear studies (outer circle) compared to the percentage of studies of each type (inner circle).

The total number of citations was 488 within the network and 9160 on Web of Science. Positive studies, comprising 33% of all studies, received 236 (48%) within-network citations and 6187 (68%) Web of Science citations. Negative studies (52% of all studies) received 205 (42%) within-network citations and 2113 (23%) Web of Science citations,

while unclear studies (15% of all studies) received 47 (10%) within-network citations and 860 (9%) Web of Science citations. The study by Caspi and colleagues (2003) [304] received a large share of the citations to positive studies, particularly in Web of Science. However, even after exclusion of this study, positive studies still received 40% of within-network and 45% of Web of Science citations, as compared to 48% and 39%, respectively, for negative studies.

On average, negative studies received 5.5 (standard deviation (SD)=9.3) within-network citations, while unclear studies received 4.3 (SD=6.2) and positive studies received 9.8 (SD=14.6). Positive studies other than Caspi and colleagues (2003) [304] received 7.4 (SD=8.9) within-network citations on average. For Web of Science, negative studies received, on average, 55.6 (SD=72.3) citations, while unclear studies received 78.2 (SD=61.6) and positive studies received 257.8 (SD=765.5) citations. Positive studies other than Caspi and colleagues (2003) received 103.8 (SD=132.4) citations on average.

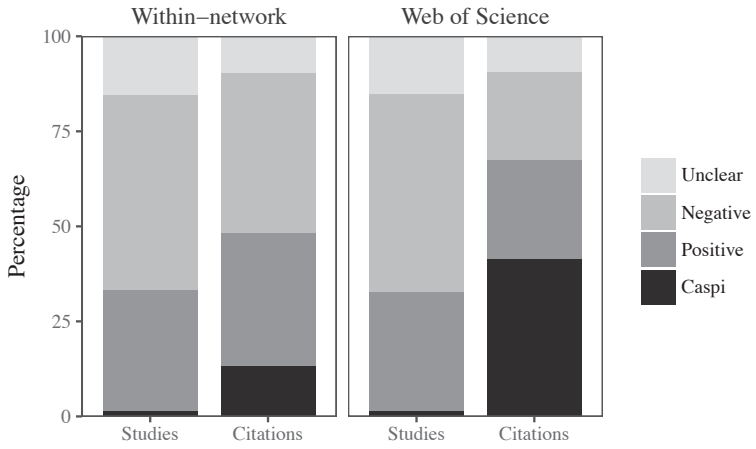

Figure 6.1: *Percentage of studies of each type (the study by Caspi and colleagues (2003), other positive studies, negative studies, and unclear studies) and the percentage of citations received by each type of study for within-network citations and Web of Science citations.*

Presence of positive focus in abstracts

Figure 6.2 depicts the presence of a positive focus within the set of studies. Of the 24 positive studies, 21 (88%) abstracts were positive and 3 (13%) abstracts were partially supportive. These partially supportive abstracts focused on gender differences (2 abstracts) or on a three-way interaction (1 abstract). Of the 11 unclear studies, 5 (45%) abstracts were partially supportive and 6 (55%) abstracts were positive. Of the 38 negative studies, 16 (42%) abstracts were negative, 9 (24%) abstracts were partially supportive, and 13 (34%) abstracts were positive. Thus, 22 out of 38 (58%) negative studies had a (partially) positive focus.

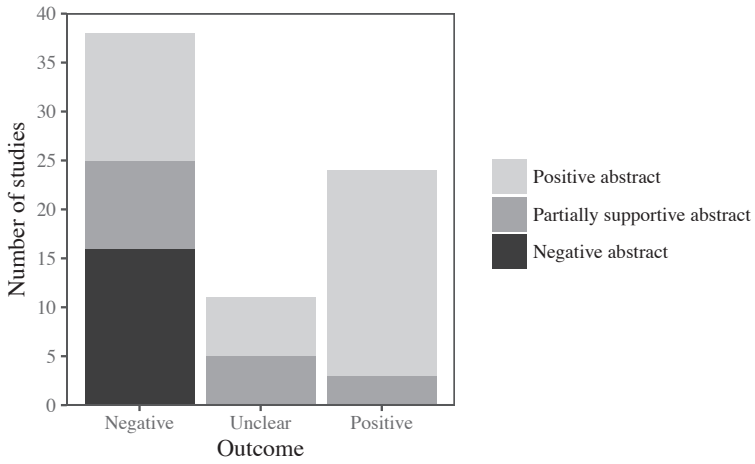

Figure 6.2: *Abstract coding by study outcome. The categories on the x-axis represent the outcome of the study, while the different sections of the bars indicate the abstract coding.*

Effect of focus on citation

Figure 6.3 shows the distribution of citations (outer circle) by presence of a positive focus (inner circle) in negative studies. Studies without a positive focus, which comprised 42% of all negative studies, received 97 (47%) out of 205 within-network citations and 679 (32%) out of 2113 Web of Science citations to negative studies. Studies with a partially positive focus (24% of all studies) received 28 (14%) within-network citations and 366 (17%) Web of Science citations, while studies with a positive focus (34% of all studies) received 80 (39%) within-network citations and 1068 (51%) Web of Science citations.

On average, a negative study without a positive focus received 6.1 (SD=9.5) citations within the network, while a study with a partially positive focus received 3.1 (SD=5.9) citations and a study with a positive focus received 6.7 (SD=11.3) citations. For Web of Science, a study without a positive focus received 42.4 (SD=44.8) citations on average, while a study with a partially positive focus received 40.7 (SD=40.5) citations and a study with a positive focus received 82.2 (SD=106.6) citations.

Citation patterns by study category

Within the network, both positive and negative studies showed preferential citation of positive studies. Although only 33% of all studies were positive, 55% of citations made by positive studies were to other positive studies, as were 45% of citations made by negative studies. Only negative studies without a positive focus (22% of all studies) additionally showed increased citation of other negative studies without a positive focus, allocating 30% of citations to these studies.

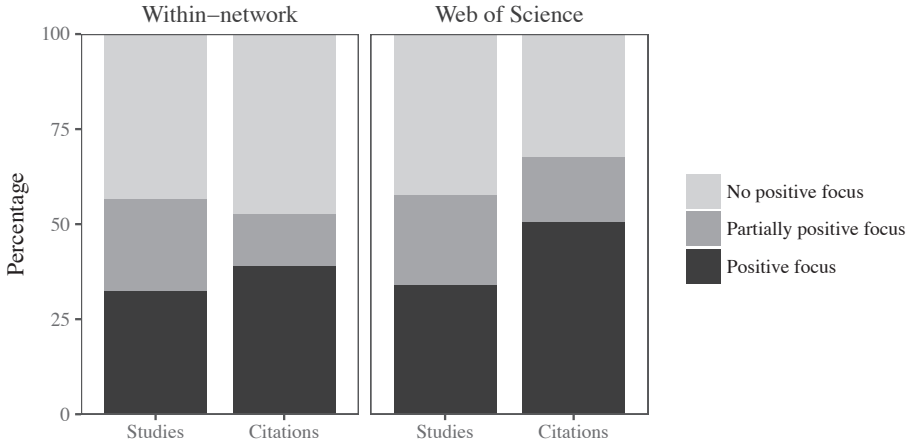

Figure 6.3: *Percentage of studies of each type (positive focus, partially positive focus, and no positive focus) and the percentage of citations received by each type of study for within-network citations and Web of Science citations.*

Sensitivity analyses

Analyses examining the percentage of subsequent studies citing a study (within-network), the yearly Web of Science citation rate, or the median number of citations (rather than the mean) yielded similar results as our main analyses (data not shown).

When we recoded studies based upon the smallest p-value rather than the average p-value, 10 negative studies and 2 unclear studies became positive. Of the smallest p-values from these 12 studies, 2 were between 0.04 and 0.05, 5 were between 0.01 and 0.05, 4 were less than 0.01, and 1 was only given as <0.05. After recoding, 36 studies were positive, 28 studies were negative, and 9 studies were unclear. The prevalence of a (partially) positive focus in the remaining negative studies decreased from 58% to 43% (12 out of 28). Recoding did not markedly affect citation patterns (data not shown).

Discussion

Principal findings

We examined citation patterns within the literature on 5-HTTLPR, life stress, and depression. In line with previous research [45, 313], we found that positive studies received more citations than negative studies. This effect was present both within the network of primary studies and within the broader literature (as represented by Web of Science citations), but it was more pronounced within the broader literature. This more pronounced

difference appeared to be largely driven by the study of Caspi and colleagues (2003) [304], which was cited especially frequently, illustrating how such a premier finding may continue to exert considerable influence even as other studies accumulate. Excluding this study reduced, but did not eliminate, citation differences between positive and negative studies.

Furthermore, we found that a (partially) positive focus was present in the abstract of over half of the negative studies. Consequently, although the majority of studies (52%) were negative, these appeared to form a fairly small minority (22%), judging by the abstracts. A positive focus did not affect citation rates within the network, but it increased citation rates within the broader literature.

This suggests that authors of other primary studies are not affected by a positive focus in abstracts. However, upon examining within-network citations to negative studies, we found that studies without a positive focus were overwhelmingly cited as negative (95%), while studies with a positive focus were usually cited as positive (56%) or partially supportive (38%), and only rarely as negative (6%). Thus, the positive focus was still propagated through these citations. Studies with a partially positive focus were actually cited less frequently than studies without a positive focus, particularly within the network. This may be because these studies, which often focused on three-way interactions, appear less relevant to the authors of primary studies on the two-way interaction itself.

Our results resemble those found previously for the literature on 5-HTTLPR and amygdala activation [315], although citation bias toward positive studies and in particular positive abstracts was more pronounced in the amygdala activation literature. This difference may be due to the controversy surrounding gene-environment interactions: both opponents and proponents may be more likely to cite negative studies when there is controversy, the former to cast doubt upon the value of gene-environment research, the latter to point out potential flaws in these negative studies. However, when we examined early citations (prior to 2010) and late citations separately, there was little evidence that citation bias toward positive studies has changed since the publication of critical meta-analyses in 2009 [307, 308], although there did seem to be a decrease in citation bias toward negative studies with a positive focus.

In this study, we extended the concept of spin, which originated within the clinical trial literature, to observational studies. Given the differences between observational studies and clinical trials, we use the term "positive focus" instead of spin. Unlike clinical trials, which are usually narrowly focused on the efficacy of an intervention, observational studies tend to have more wide-ranging topics and often lack a clearly defined, a priori primary outcome.

In this study, we specifically examined whether abstracts suggested that the results supported the 5-HTTLPR, life stress, and depression hypothesis, although some studies had other (primary) hypotheses (e.g., three-way interactions). However, all studies were

clearly inspired by Caspi and colleagues (2003) [304] and have a bearing on the original finding. As discussed by Kapur and colleagues (2012) [67], novel findings in biological psychiatry often become surrounded by a penumbra of subsequent studies with a multiplicity of measures and significant findings that are, at best, "approximate replications". A finding thus appears to be supported, even though it has not been decisively replicated (or refuted) and even though some supportive findings may have been accompanied by negative findings on a more precise replication of the original finding. We therefore deemed it important to specifically investigate how papers report on their findings with respect to the original finding by Caspi and colleagues (2003) [304].

Duncan and Keller (2011) have previously shown that negative replications of G×E findings were often published alongside positive findings [309]. This tendency, which is distinct from, although related to a focus on positive findings in the abstract, further illustrates that authors are inclined to present a positive message. The tendency for the hypothesis to expand, as reflected in the study of three- or even four-way interactions between 5-HTTLPR, life stress, and gender, other genes or environmental factors, may also be rooted, in part, in the search for positive findings.

There is a consensus that negative results are difficult to publish, which is supported by the finding that the sample size of purely negative G×E studies was six times greater than that of positive studies [309]. Although cohort studies have not found a greater journal acceptance rate for positive papers compared to negative papers [320], these studies often examined high-impact general medical journals, and authors may not submit negative studies that they judge to have little chance of acceptance to such journals. The perception that negative studies are unpublishable, as well as the conviction that the effect is real, may lead researchers to use motivated reasoning to justify presenting their findings in a positive light (without necessarily any conscious intentions of doing so) [321].

Strengths and limitations

One of the strengths of our study is our examination of positive focus in abstracts and its influence on citation patterns, as the decision to cite a study and the manner of citation may be based on the abstract only. An additional strength is that we examined citations within the network of primary studies as well as in the broader literature, since authors of other primary studies are likely to have different citation motives than authors writing on a broader or different topic. We also corrected for differences in opportunity to be cited by looking at yearly rates and the percentage of studies citing a given study, which yielded similar results.

Finally, we performed a sensitivity analysis based upon the smallest p-values, when studies had multiple relevant stressors or outcomes. Using only the smallest p-value accounts for studies in which the analysis considered most important by the authors is statistically significant, whereas other analyses are not. This lenient approach does not account for

multiple testing, although many p-values were not highly significant (only 4 out of 12 were smaller than 0.01). While this approach increased the proportion of positive studies, 43% of the remaining negative studies still had a (partially) positive focus in the abstract, and citation patterns were comparable, showing that the overall pattern remains the same even as some individual studies shift categories.

A limitation of our study is that the inter-rater agreement for coding study outcomes was only moderate. Although some disagreements were easily resolved, others reflect the opacity of some of the studies we included, which often included a multitude of stressors, outcomes, analyses, and p-values. Unfortunately, the G×E field is characterized by a proliferation of approaches, hampering easy interpretability and comparability. Pre-specification of a primary outcome and analytical approach, such as proposed in the protocol of a collaborative meta-analysis [322], may help curb this proliferation and yield clear results.

A second limitation is that we did not incorporate meta-analyses, although citations are probably diverted from primary studies to meta-analyses once these are published. However, both the negative and positive meta-analyses in this field [306, 307, 308] have been highly cited, suggesting that inclusion of meta-analyses would not undo the preferential citation of positive studies. Finally, we did not assess study quality. Arguably, high-quality studies should receive more citations, and it is possible, although not very likely [309], that positive studies were of higher quality than negative studies.

Conclusions

Although we have examined a specific, highly prominent finding, selective focus on positive findings and citation bias are unlikely to be isolated problems, limited to this particular example. On the contrary, like other biases, they are probably widespread in many scientific disciplines. Our research therefore illustrates evidence-base-distorting mechanisms that may be at work in other areas as well.

Consequently, our findings have broad implications. The frequent presence of positive conclusions in the abstracts of negative studies suggests that readers should endeavor to read the full study and personally assess its results whenever possible. Furthermore, researchers are well-advised to perform an independent search to obtain all relevant studies, as combing through reference lists may yield a disproportionate number of positive studies. Researchers should also be encouraged to cite all relevant studies, and peer reviewers may play a part in ensuring that relevant negative studies are cited and that abstracts provide an accurate and complete representation of the results.

Our study is not a meta-analysis and is not intended to provide a definitive answer to the question of whether 5-HTTLPR moderates the association between life stress and the development of depression. Instead, we examined whether there is a tendency

within this literature to preferentially cite some studies over others. We have shown that positive studies receive a disproportionate amount of attention and that negative studies are frequently presented as positive, which distorts the apparent evidence base. In the G×E field, where individual studies often include a variety of analyses and p-values, it is difficult for any reader to tell the forest from the trees. The presence of a selective focus on positive findings and citation bias further compounds this difficulty by hiding published negative results from view and rendering the "forest" greener than it truly is.

Appendix

Table 6.1: *Supplemental table of studies*

Study	Stressor	Depression measure	P-value	Outcome	Abstract
Aguilera (2009)	Child sexual abuse	SCL-90-R subscale	<0.0001	P	P
Antypa (2010)	Child emotional abuse	MDQ	0.59	N	N
Araya (2009)	SLEs	SDQ	0.23	N	N
Aslund (2009)	Childhood maltreatment	DSRS (symptoms) DSRS (diagnosis)	0.016 0.015	P	PS
Beaver (2012)	Perceived stress	CES-D	0.001	P	PS
Benjet (2010)	Relational peer victimization	CDI	0.03	P	P
Brown (2013)	Childhood maltreatment Life events	SCAN (prospective) SCAN (retrospective) SCAN	0.3916 0.0017 0.9745	N	N
Brummett (2008)	Caregiving stress Father's low education	CES-D ODS	n/a n/a	U	PS
Bull (2009) [E]	Interferon-α treatment	BDI/ZSDS	0.03	P	P
Carli (2011)	Childhood abuse	HDRS	0.0026 (O)	N	N
Caspi (2003)	SLEs	Depression symptoms Depression diagnosis Informant report	0.02 0.056 <0.01	P	P
	Childhood maltreatment	Depression diagnosis	0.05		
Cervilla (2006)	Threatening life events	CIDI	0.04	P	P
Chipman (2007)	SLEs Childhood adversity	GoDS	0.584 0.613	N	N
	Family stress Persistent family adversity	SMFQ	0.903, 0.812 0.215, 0.007 (O)		
Chorbov	Traumatic	C-SSAGA	<0.0001 (O)	N	N

continued

Table 6.1: *Supplemental table of studies*

Study	Stressor	Depression measure	P-value	Outcome	Abstract
(2007)	life events				
Cicchetti (2007)	Childhood maltreatment	DISC YSR	NS NS	N	P
Cicchetti (2011)	Childhood maltreatment	CDI & TRF	0.276	N	PS
Comasco (2011a)	Season of delivery	EPDS	n/a	U	PS
Comasco (2011b)	SLEs	EPDS	n/a	U	P
Conway (2010)	Chronic family stress	BDI SCID	NS	N	N
Coventry (2010)	SLEs	SSAGA DSSI	NS NS or <0.05 (O)	N	N
Dick (2007)	SLEs	SSAGA	n/a	U	P
Eley (2004)	Environmental risk	SMFQ	0.09	N	P
Fergusson (2011)	Childhood adversity SLEs Adult adversity	CIDI	NS or <0.05 (O) NS NS	N	N
Gibb (2009)	Maternal criticism	CDI	n/a	U	PS
Gillespie (2005)	Personal SLEs Network SLEs Personal SLEs Network SLEs	SSAGA SCL-90 + DSSI SSAGA SCL-90 + DSSI	0.43 0.66 0.74 0.15	N	N
Goldman (2010)	Lifetime trauma Major life events	CES-D	0.025 0.294	N	PS
Grabe (2005)	Unemployment Chronic disease	Von Zerssen scale	n/a n/a	U	PS
Grabe (2012)	Child abuse Adult trauma	BDI	0.2218, 0.1680 0.4030, 0.3858	N	PS
Grassi (2010)	SLEs	HAD-D	0.18	N	N
Hammen (2010)	Acute life events	BDI	NS	U	PS

continued

Table 6.1: *Supplemental table of studies*

Study	Stressor	Depression measure	P-value	Outcome	Abstract
	Chronic family stress				
Hankin (2011)	Negative life events	CDI	>0.05	N	P
Jacobs (2006)	Life events	SCL-90 subscale	0.04	P	P
Jenness (2011)	Episodic stressors	CDI	0.02	N	PS
	Chronic family stress		0.88		
Kaufman (2004)	Childhood maltreatment	MFQ	0.01	P	P
Kaufman (2006)	Childhood maltreatment	MFQ	<0.03	P	PS
Kendler (2005)	SLEs	Depression diagnosis	0.04	P	P
Kilpatrick (2007)	Hurricane exposure	SCID-IV	n/a	U	P
Kim (2007)	SLEs	GMS (B3)	n/a	U	P
Kim (2009)	Somatic disorders	GMS (B3)	0.048	P	P
Kohen (2008) [E]	Stroke	GeDS	0.045	P	P
Kraus (2007) [E]	Interferon-α treatment	HAD-D	0.413	N	N
Kumsta (2010)	Institutional deprivation	Rutter scales, SDQ, CAPA	0.14	N	P
Laucht (2009)	SLEs Family adversity	SCID BDI	p<0.05 (O) NS	N	N
Lazary (2008)	Threatening life events	ZSDS	0.0049	P	P
Lenze (2005) [E]	Hip fracture	PRIME-MD HDRS	0.026 p<0.001	P	P
Lotrich (2007) [E]	Interferon-α treatment	BDI SCID	NS p<0.05	U	P
Mitchell	Low education	CIDI-SF	NS	N	P

continued

Table 6.1: *Supplemental table of studies*

Study	Stressor	Depression measure	P-value	Outcome	Abstract
(2011)					
Mossner (2001) [E]	Parkinson's disease	HDRS	0.02	P	P
Nakatani (2005) [E]	Acute myocardial infarction	ZSDS	0.01	P	P
Otte (2007) [E]	Coronary disease	DIS	0.04	P	P
Petersen (2012)	SLEs	YSR + CBCL	0.035	P	P
Phillips-Bute (2008) [E]	Coronary artery bypass surgery	CES-D	0.70, 0.02	N	N
Power (2010)	Recent life events	MINI CES-D	0.08, 0.70 0.47, 0.82	N	N
Quinn (2012)	Early life stress	MINI	NS	N	PS
Ramasubbu (2006) [E]	Stroke	SCID	0.025	P	P
Ressler (2010)	Childhood trauma	BDI SCID	NS 0.016	N	P
Ritchie (2009)	Excessive problem sharing by parents Poverty Other childhood trauma	MINI + CES-D + antidepressant use (composite)	0.027 (O) 0.025 (O) NS	N	PS
Scheid (2007)	Life stressors	CES-D	>0.24 (×6), 0.04	N	P
Scheid (2011)	Life stressors	CES-D	NS	N	PS
Sen (2010) [E]	Medical internship	PHQ-9	0.002	P	P
Sjöberg (2006)	Residence type Parental separation Traumatic family conflict Composite	DSRS	0.106 0.106 0.01 0.004	N	PS
Stefanis	Military	SCL-90-R	0.11	N	P

continued

Table 6.1: *Supplemental table of studies*

Study	Stressor	Depression measure	P-value	Outcome	Abstract
(2011) [E]	conscription				
Sugden (2010)	Bullying	CBCL + TRF	0.013, 0.508	N	P
Surtees (2006)	Adverse life events	HLEQ	NS	N	N
Taylor (2006)	Early life stress	BDI	<0.008	P	P
	Negative life events		<0.024		
Tsuboi (2011)	Perceived stress	CES-D	<0.05	P	P
Uher (2011)	Childhood maltreatment	DIS	0.003, 0.2312	N	P
Wichers (2008)	Childhood trauma	SCL-90-R subscale SCID	0.4 0.4	N	PS
Wilhelm (2006)	Adverse life events	DIS & CIDI	0.036, 0.435	N	P
Wilhelm (2012) [E]	Diabetes	Clinical interview PHQ-9 K10	0.81 NS 0.047	N	P
Zalsman (2006)	SLEs	HDRS BDI	0.04 0.51	N	P
	Child abuse	HDRS BDI	0.20 0.19		
Zhang (2009a)	Negative life events	DIGS	0.005 - 0.006 (O)	N	P
Zhang (2009b) [E]	Parkinson's disease	CES-D	0.804	N	N

[E] indicates a study with an exposed-only sample.

List of acronyms for stressors and depression measures: *BDI: Beck Depression Inventory; CAPA: Child and Adolescent Psychiatric Assessment; CBCL: Child Behavior Checklist; CDI: Children's Depression Inventory; CES-D: Center for Epidemiological Studies Depression Scale; CIDI: Composite International Diagnostic Interview; CIDI-SF: Composite International Diagnostic Interview – Short Form; C-SSAGA: Child Semi-Structured Assessment for the Genetics of Alcoholism; DIGS: Diagnostic Interview for Genetic Studies; DIS: Diagnostic Interview Schedule; DISC: Diagnostic Interview Schedule for Children; DSRS: Depression Self-Rating Scale; DSSI: Delusions-Symptoms-States Inventory; EPDS: Edinburgh Postnatal Depression Scale; GeDS: Geriatric Depression Scale; GoDS: Goldberg Depression Scale; GMS: Geriatric*

Mental State schedule; HAD-D: Hospital Anxiety and Depression scale – depression subscale; HDRS: Hamilton Depression Rating Scale; HLEQ: Health and Life Experiences Questionnaire; K10: Kessler Psychological Distress Scale; MDQ: Major Depression Questionnaire; MFQ: Mood and Feelings Questionnaire; MINI: Mini International Neuropsychiatric Interview; ODS: Obvious Depression Scale; PHQ-9: Patient Health Questionnaire; PRIME-MD: Primary Care Evaluation of Mental Disorders; SCAN: Schedules for Clinical Assessment in Neuropsychiatry; SCID: Structured Clinical Interview for DSM Disorders; SCL-90-R: Symptom Checklist 90 Revised; SDQ: Strengths and Difficulties Questionnaire; SLEs: stressful life events; SMFQ: Short Mood and Feelings Questionnaire; SSAGA: Semi-Structured Assessment for the Genetics of Alcoholism; TRF: Teacher Report Form; YSR: Youth Self-Report; ZSDS: Zung Self-rating Depression Scale.

P-values: *P-values apply to the specific combination of stressor and depression measure listed. "NS" stands for non-significant, in cases where an exact p-value was unavailable. n/a indicates those studies for which the p-value for a two-way interaction could not be found and which were subsequently coded as unclear. Multiple p-values may be given for a single stressor-depression measure combination in case of, for example, multiple time points. For significant p-values, (O) indicates a significant interaction in the opposite direction from that expected. For non-significant p-values, the direction of the effect (expected or opposite) is not indicated.*

Outcome and abstract coding acronyms: *P: Positive; N: Negative; U: Unclear; PS: Partially supportive.*

Chapter 7

The cumulative effect of reporting and citation biases on the apparent efficacy of treatments: the case of depression

Ymkje Anna de Vries, Annelieke M. Roest, Peter de Jonge,
Pim Cuijpers, Marcus R. Munafò, Jojanneke A. Bastiaansen

Submitted

Abstract

Study publication bias is recognized as an important threat to evidence-based medicine, but other biases also affect the quality of the evidence base. Using the evidence base for antidepressants and psychotherapy, we illustrate how the effects of study publication bias, outcome reporting bias, spin, and citation bias accumulate to hide negative results from view.

Within the antidepressant literature, 52 (50%) of 105 trials were negative, but only 4 (5%) published trials unambiguously reported that the treatment was not effective. Other negative trials remained unpublished (27 trials), were published as if positive due to outcome reporting bias (10 trials), or were published with spin in the abstract or discussion (11 trials). Compounding the problem, trials reporting positive results were cited, on average, three times as frequently as negative trials (92 versus 32 citations).

Within the psychotherapy literature, we obtained 142 published trials. Although 49 (35%) of these trials were considered to be negative, only 12 (8%) abstracts concluded that psychotherapy was not more effective than a control condition. Positive psychotherapy trials were also cited more frequently than negative trials (111 citations versus 58).

These results show the pernicious cumulative effect of reporting and citation biases. Even when the decision to publish a trial is made, there are still several hurdles to pass before negative results can receive the visibility they deserve. Mandatory universal registration, in combination with openness to negative results and vigilance on the part of peer reviewers, journal editors, and readers, may help to prevent and uncover bias.

Introduction

Evidence-based medicine is the cornerstone of good clinical practice, but it is dependent on the quality of the evidence upon which it is based [323]. It is well-known that trials with statistically significant findings are more likely to be published than those with non-significant findings; this is a major problem since as many as half of all randomized controlled trials (RCTs) are never published [38].

However, negative trials face additional hurdles beyond study publication bias that can result in the disappearance of non-significant results [38, 40, 44]. Here, we analyze the cumulative impact of biases on the apparent efficacy of treatments, and discuss possible remedies, using the evidence base for two effective treatments for depression: antidepressants and psychotherapy [19, 324].

Reporting and citation biases

Many different biases exist and can affect an evidence base. We distinguish between four major biases: study publication bias, outcome reporting bias, spin, and citation bias. While study publication bias involves non-publication of an entire study, outcome reporting bias refers to non-publication of negative outcomes or non-significant analyses within a published article [38]. It also refers to relegating non-significant primary outcomes to a secondary status or upgrading a protocol-specified secondary (but statistically significant) outcome to a primary outcome in the publication. Both study publication bias and outcome reporting bias can be an important threat to the validity of meta-analytic conclusions regarding treatment efficacy [102, 325].

Although trials that faithfully report non-significant results on the primary outcome will yield accurate effect estimates for use in meta-analyses, the interpretation of results can still be positively biased, which may affect the apparent efficacy of treatments. The use of specific reporting strategies that could distort the interpretation of results and mislead readers is defined as spin [40]. Spin occurs, for instance, when authors focus on statistically significant results in secondary analyses and conclude that the treatment is effective despite non-significant results on the primary outcome. It has been shown that a treatment is rated as being more beneficial when the abstract of a journal article has been spun [41].

Finally, even when a trial with non-significant results is published and accurately reported, citation bias is an additional obstacle to ensuring that it receives as much attention as a study with significant results. Studies with positive results receive more citations than negative studies in many medical fields, including psychiatry [45, 313], which results in a heightened visibility of positive results.

The evidence base for antidepressants

To examine the cumulative effect of these biases, we assembled a cohort of 105 trials of antidepressants for depression. Seventy-four of these trials were included in the 2008 study by Turner et al. [19], to which we added 31 trials from the Food and Drug Administration (FDA) database [105] used to support the marketing approvals of four novel antidepressants after the Turner et al. study. Pharmaceutical companies are required to preregister all trials that they intend to use in support of such an application with the FDA; hence, trials with non-significant results may not be published but are still accessible.

Figure 7.1 demonstrates the cumulative impact of reporting and citation biases on these trials. In the initial cohort of 105 antidepressant trials, 53 (50%) trials were considered to be positive by the FDA and 52 (50%) were considered negative or questionable (Figure 7.1a). While all but one of the positive trials was published (98%), only 25 (48%) of the negative trials were published. Hence, the initial cohort is reduced to 77 published trials, of which 25 (32%) were negative according to the FDA (Figure 7.1c).

Ten of these published negative trials, however, became 'positive' trials in the published literature, as a consequence of omitting unfavorable outcomes or switching the status of primary and secondary outcomes (Figure 7.1c). Without access to the FDA reviews, it would not have been possible to tell that these trials, when analyzed according to the pre-specified protocol, were not positive.

Among the remaining 15 (19%) negative trials, five were published (in four articles) with spin in the abstract (i.e., concluding that the treatment was effective). Five additional articles contained mild spin in the abstract (e.g., suggesting that the treatment is at least numerically better than placebo, or that the results cannot be interpreted because the trial, rather than the drug, failed, since an active comparator was not more effective than placebo either) [326]. One article did not have an abstract, but the discussion section concluded that there was a "trend for efficacy". Hence, only four (5%) of 77 published trials unambiguously reported that the treatment was not effective (Figure 7.1d).

Compounding the problem, trials reporting positive results were cited three times as frequently as negative trials (92 versus 32 citations in Web of Science, January 2016) (Figure 7.1e). Among negative trials, those with (mild) spin in the abstract received an average of 36 citations, while those with a clearly negative abstract received 25 citations on average. Although there were too few published negative antidepressant trials for a definitive conclusion, this suggests a possible synergistic effect between spin and citation bias, where negatively presented negative trials receive especially few citations [315, 327]. Altogether, these results show that the effects of different biases accumulate to hide non-significant results from view.

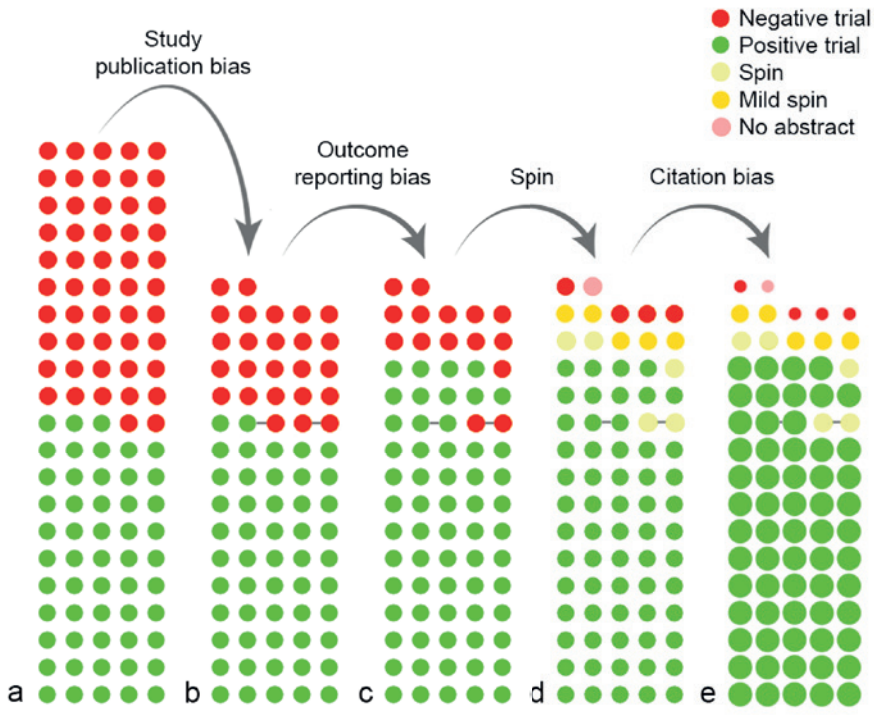

Figure 7.1: *Panel a displays the original, complete cohort of trials, while Panels b through e show the cumulative effect of biases. Each circle indicates a trial, while the color indicates the results (red for negative, green for positive) or the presence of spin (yellow and orange for strong and mild spin, respectively, and pink for a trial that was published in an article without an abstract). Circles connected by a grey line indicate trials that were published together in a pooled publication. In Panel e, the size of the circle indicates the (relative) number of citations received by that category of studies.*

The evidence base for psychotherapy

While the pharmaceutical industry has a financial motive for suppressing unfavorable results, these biases are also present in other areas of research, such as psychotherapy. In the absence of a standardized registry of trials, however, their presence is more difficult to assess, and the various biases are more difficult to disentangle. Statistical tests suggest an excess of positive findings in the psychotherapy literature, which may be due to either study publication bias or outcome reporting bias [177, 328].

More recently, in a cohort of psychotherapy trials funded by the National Institutes of Health, which maintains a database of funded grants and hence functions as a (limited) registry, it was found that 13 (24%) out of 55 initiated trials were never published [35]. The effect size of these unpublished studies was markedly lower than that of published

studies, suggesting a bias against publishing negative findings.

Regarding spin in the psychotherapy literature, we found that 49 (35%) of 142 papers were considered to be negative in a meta-analysis by Flint et al. [328], but only 12 (8%) abstracts concluded that psychotherapy was not more effective than a control condition. The remaining 130 abstracts reported either positive (73%) or mixed findings (19%), concluding, for example, that the treatment was effective for one outcome but not another.

Although we could not establish the pre-specified primary outcome for these trials, and therefore cannot determine whether a specific abstract is biased, it is clear that published psychotherapy trials, as a whole, provide an impression of the effectiveness of psychotherapy that is more positive than justified by the available evidence. Similar to the antidepressant literature, positive psychotherapy trials were cited more frequently than negative trials (111 citations versus 58). Compounding the problem, negative trials with a positive or mixed abstract received an average of 59 and 87 citations, respectively, while those with a negative abstract received only 26 citations.

Preventing bias

Mandatory and universal prospective registration has long been advocated as a solution to the widespread problem of study publication bias [329]. It could also be an effective solution to the problem of outcome reporting bias. Since 2005, the International Committee of Medical Journal Editors (ICMJE) has required prospective registration of clinical trials as a precondition for publication [330], but many journals do not require registration [331] and others allow retrospective registration [332]. Since 2007, investigators are also legally required by the FDA to prospectively register most phase 2 and 3 drug trials (i.e., clinical trials investigating efficacy and safety in patients).

The increasing pressure to publicly register trials may explain why negative trials of novel antidepressants are more frequently published than those of antidepressants approved earlier. In particular, all unpublished negative or questionable trials of novel antidepressants were completed before 2004, while the 25 trials completed in 2004 or later (including 14 trials for which registration was legally required) were all published, even though nine of these trials were considered negative or questionable.

A regulatory requirement is likely to be one of the most effective measures to ensure universal registration. It is therefore unfortunate that the 2007 law is not fully comprehensive, as it excludes trials of behavioral interventions (such as psychotherapy) and phase 1 (healthy volunteer) trials.

However, registration in and of itself also seems to be insufficient to ensure complete and accurate reporting of a trial. It has been found that only around half of all trials registered in ClinicalTrials.gov were published within two years of completion (as required by the

FDA Amendments Act) [333]. The COMPare Project (compare-trials.org), which aims to track outcome switching, and other studies [334] found that non-reporting of outcomes and the addition of novel outcomes (without any indication that these outcomes had not been pre-specified) was also common.

Hence, close examination of clinical trial registries by independent researchers may be necessary for trial registration to be a truly effective deterrent to study publication bias and outcome reporting bias. For already completed drug trials, consulting FDA reviews as we have done for this analysis may also be helpful. An alternative (or addition) to registration could be the publication of study protocols or "registered reports", in which journals accept a study for publication on the basis of the introduction and methods and before the results are known.

It may prove to be more difficult to prevent spin and citation bias. Peer reviewers could play a crucial role in ensuring that abstracts accurately report a trial's findings and that important negative results are cited. Journals also have a role to play in ensuring that researchers do not feel like they must 'oversell' the significance of their results to get published.

The prevalence of spin and citation bias also indicates the importance of assessing a study's actual results (rather than relying on the authors' conclusions) and of conducting an independent literature search whenever possible, since the reference lists of other papers may yield a disproportionate number of positive (and positively presented) studies.

Conclusions

The problem of study publication bias, which was first recognized over 50 years ago [36], is now well-known. Our examination of antidepressant trials, however, shows the pernicious cumulative effect of additional reporting and citation biases. Even when the decision to publish a trial has been made, there are still several hurdles to pass before negative results can receive as much visibility as positive results.

Within the antidepressant literature, we found that the cumulative impact of these biases eliminates the majority of negative results from the published literature and leaves the few remaining negative results that do get published much more difficult to discover than positive results. These biases are unlikely to be unique to antidepressant trials, which we have taken as a case study. We have shown that similar processes, though more difficult to assess, are at work within the psychotherapy literature.

The presence of these individual biases has been shown in various medical fields [38, 40, 313], and it is likely that their effects also accumulate whenever these biases are present. Consequently, researchers and clinicians across medical fields must be aware of the potential for bias to distort the apparent efficacy of an intervention.

Chapter 8

Poor adherence to guidelines for antidepressant initiation in children and adolescents in the Netherlands

Ymkje Anna de Vries, Peter de Jonge, Luuk Kalverdijk, Jens H. J. Bos, Catharina C. M. Schuiling-Veninga, Eelko Hak

European Child & Adolescent Psychiatry (2016), 25, 1161 - 1170

Abstract

Background: The Dutch guideline for the treatment of depression in young people recommends initiating antidepressant treatment with fluoxetine, as the evidence for its efficacy is strongest and the risk of suicidality may be lower than with other antidepressants. Furthermore, low starting doses are recommended. We aimed to determine whether antidepressant prescriptions are in accord with guidelines.

Methods: A cohort of young people aged between 6 and 17 at the time of antidepressant initiation was selected from IABD, a Dutch pharmacy prescription database. The percentage of prescriptions for each antidepressant was determined. Starting and maintenance doses were determined and compared with recommendations for citalopram, fluoxetine, fluvoxamine, and sertraline.

Results: During the study period, 2942 patients initiated antidepressant treatment. The proportion of these young people who were prescribed fluoxetine increased from 10.1% in 1994 – 2003 to 19.7% in 2010 – 2014. However, the most commonly prescribed antidepressants were paroxetine in 1994 – 2003 and citalopram in 2004 – 2014. The median starting and maintenance doses were ≤ 0.5 DDD/day for tricyclic antidepressants and 0.5 – 1 DDD/day for SSRIs and other antidepressants. Starting doses were guideline-concordant 58% of the time for children, 31% for preteens, and 16% for teens. Sixty percent of teens were prescribed an adult starting dose.

Conclusions: Guideline adherence was poor. Physicians preferred citalopram over fluoxetine, in contrast to the recommendations. Furthermore, although children were prescribed a low starting dose relatively frequently, teens were often prescribed an adult starting dose. These results suggest that dedicated effort may be necessary to improve guideline adherence.

Introduction

Practice guidelines in the Netherlands [335] and internationally (e.g., [336]) recommend that medication should only be prescribed to children and adolescents suffering from (moderate to) severe depression. These guidelines also recommend that pharmacotherapy should be initiated with fluoxetine, with sertraline or citalopram used in case of non-response to fluoxetine. Other antidepressants, such as mirtazapine, venlafaxine, and tricyclic antidepressants, are not recommended. In addition, treatment should be initiated with a low starting dose (a quarter to a half of the adult starting dose) [336, 337, 338].

Second-generation antidepressants may be moderately effective for depression in children and adolescents [339], but they have also been associated with an increased risk of suicidal ideation and behavior [190]. In 2004, the United States Food and Drug Administration issued a black box warning on antidepressants to emphasize the risk of suicidality in young people. There is some evidence to suggest that the risk may vary by antidepressant, with fluoxetine showing less of an increased risk than many other second-generation antidepressants [340].

The risk of suicidality may also be dose-related, with young people prescribed higher-than-modal starting doses of antidepressants showing an increased risk of suicidal behavior compared to those prescribed the modal dose [341]. Further epidemiological evidence in adults also suggests that lower-than-modal doses may be associated with decreased risk, although confounding by indication cannot be excluded. This study also found that risk was particularly increased within the first three months after starting an antidepressant [342].

Although prescription trends in children have been examined extensively [343, 344, 345, 346, 347, 348, 349, 350], most studies have not examined specifically the first prescription of an antidepressant and only one study, to our knowledge, has examined whether appropriate dosages are used [351]. This study found that antidepressant treatment in young people in the United States was more commonly initiated with a low dose after the black box warning was issued in 2004, although low doses were still only prescribed in a minority of cases.

In the Netherlands, citalopram is the most commonly prescribed antidepressant to young people [352], which suggests that the guidelines may not be followed; however, as this includes all antidepressant prescriptions and citalopram is recommended as a second antidepressant, the evidence is not yet conclusive. To our knowledge, no evidence is currently available regarding antidepressant dosing in the Netherlands.

In the current study, we therefore aimed to answer the following questions: first, do physicians initiate antidepressant treatment in young people with fluoxetine? Second, what are the usual starting and maintenance doses of antidepressants in young people, and are these in accord with the guidelines?

Methods

Data source

Prescription data were obtained from the IADB database, which contains information on prescriptions filled in community pharmacies in the Netherlands between 1994 and 2014 [353]. Patients are included in the database the first time they fill a prescription in one of the participating pharmacies. The database population in any given year is currently approximately 600.000.

The database includes information about the patient (gender, date of birth) and the prescription (fill date, Anatomical Therapeutical Chemical (ATC) code, number of tablets, daily dose [in number of tablets], and the total number of defined daily doses [DDDs] in the prescription). A DDD is defined as the assumed average maintenance dose for a drug used for its main indication in adults [354]. All outpatient prescriptions are included in the database, but inpatient prescriptions and over-the-counter medications are not.

Patient selection

From the IADB database, a cohort of young patients initiating treatment with an antidepressant was selected. We included tricyclic antidepressants (ATC-code N06AA), selective serotonin reuptake inhibitors (SSRIs, ATC-code N06AB), and other antidepressants (N06AX). Monoamine oxidase inhibitors (selective and non-selective MAOIs, ATC-codes N06AF and N06AG) were not included, as it is unlikely that a MAOI would be prescribed as the first antidepressant.

Patients were included in the cohort if they were between 6 and 17 years of age (inclusive), had been included in the database for at least six months at the time of first prescription of an antidepressant, and had not previously received a prescription for a different antidepressant. Patients aged between six and nine were categorized as children; patients aged between ten and thirteen as preteens; and patients aged between fourteen and seventeen as teens. For patients who had multiple episodes of antidepressant treatment during the study period, we only considered data from the first eligible treatment episode.

We excluded patients who were likely prescribed an antidepressant for non-psychiatric indications, specifically bed-wetting and pain. Patients starting on amitriptyline or imipramine who also received a prescription for desmopressin (H01BA02), the first-line treatment for bed-wetting, at any time during the study period were excluded (238 patients). We also excluded patients who received two or more prescriptions for pain-related medication within six months prior to initiation of a tricyclic antidepressant (111 patients). Pain-related medication was defined as any medication with ATC-code M01 (anti-inflammatory and anti-rheumatic drugs), N02 (analgesics), N03AX12 (gabapentin),

Table 8.1: *Dose equivalents*

Drug	1 DDD-equivalent (mg)
Amitriptyline	75
Citalopram	20
Clomipramine	100
Fluoxetine	20
Fluvoxamine	100
Imipramine	100
Mirtazapine	30
Paroxetine	20
Sertraline	50
Venlafaxine	100

Dose in milligrams equivalent to 1 defined daily dose (DDD) for the ten most commonly prescribed antidepressants.

and N03AX16 (pregabalin). One patient received a first prescription for two different antidepressants on the same day and was also excluded.

Data analysis

We split the data into three time periods: 1994 through 2003, 2004 through 2009, and 2010 through 2014. These time periods were chosen based upon major events: in 2004, knowledge of a possible link between antidepressants and suicidality in children became widespread, while in December 2009, the youth addendum to the Dutch Multidisciplinary Guideline for Depression was published.

We determined which antidepressant was first prescribed to each patient and calculated the percentage of patients who were prescribed fluoxetine as their first antidepressant. Possible moderators were examined by stratifying the data based upon prescriber (general practitioner (GP) or specialist) and age group (child, preteen, or teen).

We then determined the starting dose of antidepressants (in DDD/day) for each patient. A conversion of DDDs to the equivalent dose in milligrams for the ten most commonly prescribed antidepressants is provided in Table 8.1. The dose was calculated as the total number of DDDs divided by the total number of days in the first prescription. The number of days was calculated by dividing the total number of units (pills) by the number of pills to take daily.

For patients who received multiple prescriptions for the same antidepressant on the same day, we took the prescription with the lowest daily dose. Additionally, for prescriptions for the highly concentrated liquid formulations of citalopram and escitalopram, we divided the daily dose by 20 (as 1 drop of solution is approximately equivalent to 0.05

ml) before calculating the DDD/day. We excluded patients with missing daily doses (4 (0.1%) patients) or with unrealistically low (0 DDD/day, 256 (8.7%) patients) or high (>3 DDD/day, 6 (0.2%) patients) doses, as these are likely to reflect data entry errors (in particular, entering 0 as the number of units per day).

The maintenance dose of antidepressants was determined in a similar fashion. Maintenance was defined as a period in which at least 2 prescriptions with the same dose were filled, containing a minimum of 60 days' supply. Prescriptions were required to be overlapping, i.e., the number of days in the first prescription must be sufficient to cover the fill date of the subsequent prescription, after adding 25% to the number of days to account for possible non-compliance.

For patients who had multiple maintenance periods, we selected the period with the longest duration and the highest dose (if multiple periods had the same duration). Missing doses were set to 0. We excluded patients with unrealistically low (0 DDD/day, 8 (0.1%) patients) or high maintenance doses (>4 DDD/day, 0 patients). Fifty-three percent of all patients had at least one maintenance period with a realistic dose.

The distribution of starting doses and maintenance doses was determined for each antidepressant. For the SSRIs fluoxetine, citalopram, sertraline, and fluvoxamine, the distribution was compared to the Dutch dosing guidelines for children [337, 338]. These guidelines recommend a starting dose of 5 mg (0.25 DDD) for fluoxetine and citalopram, and 25 mg (0.25 DDD) for fluvoxamine. For sertraline, the guidelines both recommend 25 mg (0.5 DDD) for young children, but one guideline recommends a higher dose of 50 mg (1 DDD) for adolescents aged 13 and older [337]. We chose to compare the distribution to the latter, more lenient guideline.

Subsequently the data was stratified by prescriber and by age group to examine the possible moderating influence of these variables. We also examined how many patients were prescribed a starting dose of fluoxetine of 10 mg or less, as recommended in international guidelines (eg, United Kingdom [336]).

Results

Demographics

A total of 2942 patients were prescribed a first antidepressant during the study period and met inclusion criteria: 1194 in 1994 – 2003, 815 in 2004 – 2009, and 933 in 2010 - 2014. Of these patients, 1739 (59%) were female, 1188 (40%) were male, and for 15 (1%) information about sex was missing.

The average age of the sample at initiation was 14.2 years. Three-hundred and eleven (11%) patients were children, 573 (19%) were preteens, and 2058 (70%) were teens. The

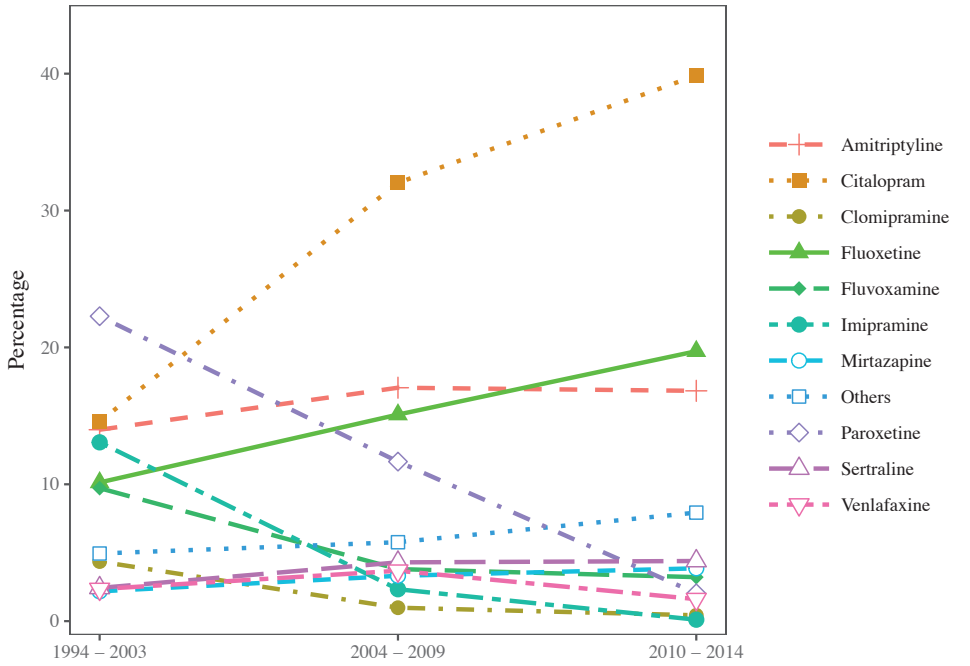

Figure 8.1: *Percentage of prescriptions for each of the ten most commonly prescribed antidepressants per time period. Prescriptions for all other antidepressants were combined into the category 'Others'.*

majority of patients (62%) received their first prescription for an antidepressant from their GP in 1994 – 2003, but by 2010 – 2014 69% of patients received their first prescription from a specialist.

First antidepressant

Of all young people initiating treatment with an antidepressant, the proportion prescribed fluoxetine increased during the study period from 10.1% in 1994 - 2003 to 19.7% in 2010 – 2014 (Figure 8.1), but fluoxetine was never the most commonly prescribed antidepressant. Instead, antidepressant treatment was most commonly initiated with paroxetine in 1994 - 2003 and with citalopram from 2004 onward. A full listing of all antidepressants is provided in Table 8.3 in the Appendix.

Stratification by prescriber (Figure 8.2) showed that specialists were slightly more likely to prescribe fluoxetine than GPs at each time point. In 2010 – 2014, GPs initiated treatment with fluoxetine in 15.8% of cases, while specialists did so in 21.5% of all cases. Specialists most commonly initiated treatment with citalopram, while GPs most commonly initiated treatment with amitriptyline. Both GPs and specialists showed a steep decrease in the

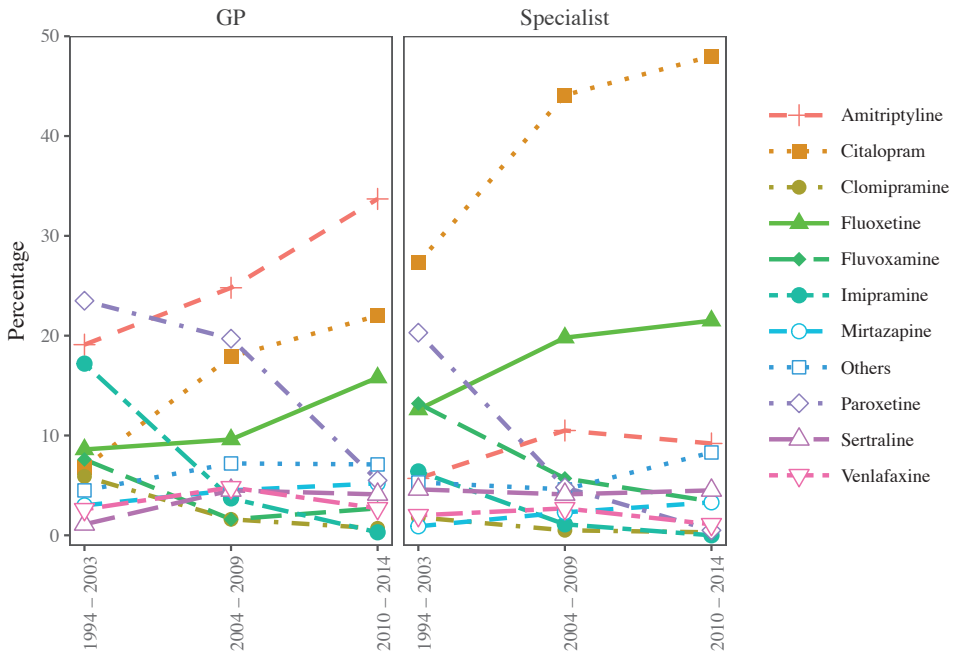

Figure 8.2: *Percentage of prescriptions for each of the ten most commonly prescribed antidepressants per time period, stratified by prescriber. Prescriptions for all other antidepressants were combined into the category 'Others'.*

use of paroxetine, although this decrease occurred earlier for specialists than for GPs.

Stratification by age group showed low rates of antidepressant initiation with fluoxetine in each age group (Figure 8.3). Children were prescribed fluoxetine least frequently (<6.5% throughout the study period). Preteens were prescribed fluoxetine in 8.6% of cases in 1994 – 2003, increasing to 15.2% in 2010 – 2014. Teens were prescribed fluoxetine relatively frequently, at 11.4% in 1994 – 2003 and 22.9% in 2010 – 2014. Citalopram was the most commonly prescribed antidepressant in all age groups by 2010 – 2014. In particular, children and preteens received citalopram in 70.9% and 56.0% of cases respectively in 2010 – 2014.

Starting doses

The median starting dose for tricyclic antidepressants was quite low, at around 0.1 – 0.3 DDD/day. For SSRIs, the median starting dose was 1 DDD/day for fluoxetine, paroxetine, escitalopram, and sertraline, while it was 0.5 DDD/day for citalopram and fluvoxamine. Mirtazapine and venlafaxine had median starting doses of 0.5 and 0.75 DDD/day respectively. A full listing of starting doses is given in Table 8.4 in the Appendix.

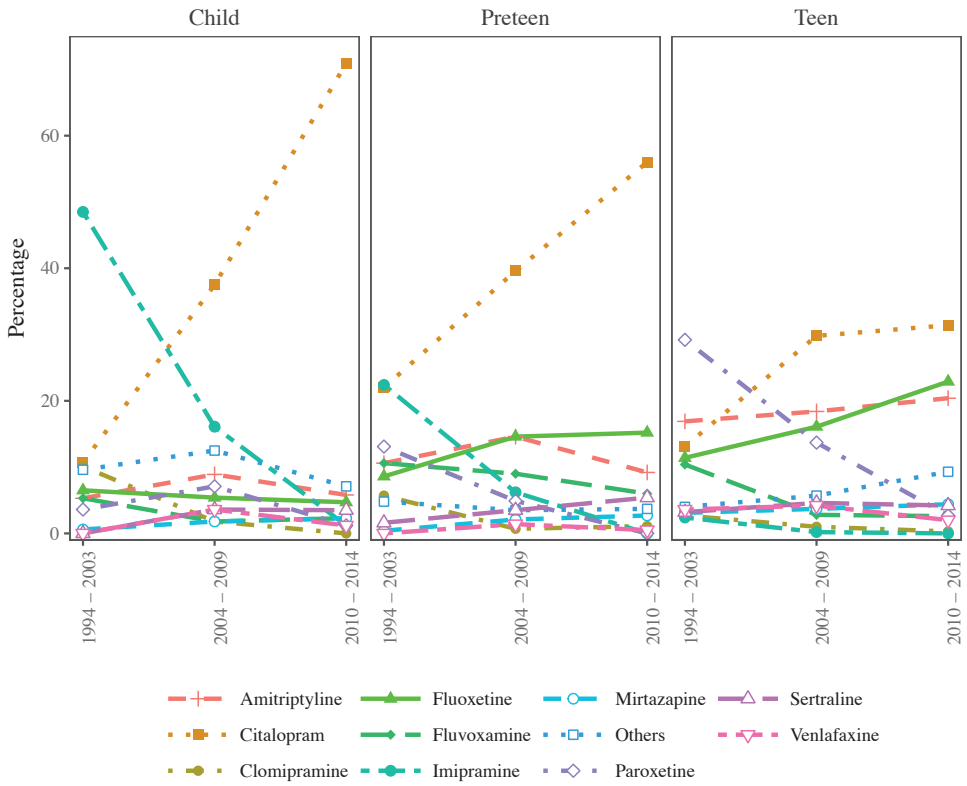

Figure 8.3: *Percentage of prescriptions for each of the ten most commonly prescribed antidepressants per time period, stratified by age group. Prescriptions for all other antidepressants were combined into the category 'Others'.*

Median starting doses were similar throughout the study period for most antidepressants, but decreased for citalopram, paroxetine, mirtazapine, and venlafaxine. For citalopram, the median starting dose decreased from 1 DDD/day in 1994 – 2003 to 0.4 DDD/day in 2010 – 2014; for paroxetine and mirtazapine, the median starting dose decreased from 1 DDD/day in 1994 – 2003 to 0.5 DDD/day in 2010 – 2014; and for venlafaxine, the median starting dose decreased from 0.75 DDD/day in 1994 – 2003 to 0.375 DDD/day in 2010 – 2014. Guideline adherence also improved somewhat within the study period for citalopram and fluvoxamine, with adherence rates being 44.3% for citalopram and 28.6% for fluvoxamine in 2010 – 2014.

Stratification by prescriber showed few differences between prescribers, although specialists prescribed some SSRIs in slightly lower doses than GPs. Stratification by age showed that children tended to be prescribed lower doses (0.5 DDD/day), particularly for the SSRIs. Preteens received slightly higher doses: 0.5 DDD/day for all SSRIs except sertraline (1 DDD/day). Teens received the highest doses, at 1 DDD/day for all SSRIs except

Table 8.2: *Guideline compliance of starting doses*

	Guideline-compliant dose			Adult dose		
	Child n (%)	Preteen n (%)	Teen n (%)	Child n (%)	Preteen n (%)	Teen n (%)
Citalopram	49 (66.2)	62 (36.3)	60 (13.6)	6 (8.1)	32 (18.7)	231 (52.4)
Fluoxetine	5 (33.3)	10 (16.1)	11 (3.4)	2 (13.3)	25 (40.3)	222 (69.2)
Fluvoxamine	4 (36.4)	6 (13.0)	14 (13.0)	0 (0.0)	8 (17.4)	50 (46.3)
Sertraline	3 (60.0)	13 (72.2)	67 (91.8)	2 (40.0)	10 (55.6)	62 (84.9)
Total	61 (58.1)	91 (30.8)	152 (15.8)	10 (9.5)	75 (25.3)	565 (59.9)

Comparison of antidepressant starting doses with guidelines and with adult doses (≥1 DDD/day). n indicates the number of prescriptions that were at or below the guideline dose.

fluvoxamine (0.5 DDD/day).

Table 8.2 shows the percentage of first prescriptions according to guidelines for fluoxetine, citalopram, sertraline, and fluvoxamine, stratified by age group. Overall, a minority of first prescriptions (22.2%) were according to guidelines: 6.5% for fluoxetine, 24.9% for citalopram, 86.5% for sertraline, and 14.5% for fluvoxamine.

Children were reasonably likely to be prescribed according to guidelines (58% across all four antidepressants), although 10% of children were prescribed an adult starting dose. On the other hand, very few teens (16%) were prescribed according to guidelines, while 60% of teens were prescribed an adult starting dose. For fluoxetine specifically, 33% of children, 16% of preteens and 3% of teens received a guideline-compliant dose (0.25 DDD/day). The corresponding percentages for a fluoxetine dose of 0.5 DDD/day (10 mg) were 67% for children, 58% for preteens, and 30% for teens.

Maintenance doses

Maintenance doses were similar to starting doses (Table 8.4 in the Appendix). The median maintenance dose for tricyclic antidepressants was around 0.2 – 0.3 DDD/day. For SSRIs, the median maintenance dose was 1 DDD/day. For the other antidepressants, median maintenance doses were 0.5 DDD/day for mirtazapine and 0.75 DDD/day for venlafaxine. Maintenance doses were nearly always according to guidelines for fluoxetine (98%), citalopram (96%), sertraline (91%), and fluvoxamine (93%).

Among those who had a valid starting dose as well as a valid maintenance dose, 60% remained at their starting dose, while 35% titrated up to a higher dose and 5% titrated down. These percentages were similar across prescribers and age groups; however, they did vary according to the antidepressant prescribed. Of the 10 most commonly prescribed antidepressants, up-titration was more likely for citalopram (46%), sertraline (42%), and

venlafaxine (47%), while it was less likely for imipramine (25%), amitriptyline (17%), paroxetine (24%), and mirtazapine (21%).

Discussion

Principal findings

Physicians initiated pharmacotherapy with fluoxetine less than 20% of the time, even after publication of the guidelines for youth in 2009. The percentage of first prescriptions for paroxetine decreased sharply after 2003, a trend which is most likely due to its particularly prominent association with suicidality in young people. Our results suggest that prescriptions for paroxetine were not replaced with fluoxetine, as the guidelines suggest, but with citalopram, which became the most popular antidepressant by 2004 – 2009.

Although citalopram is effective for depression in adults [19], it has not been shown to be effective in children and adolescents, in contrast to fluoxetine [339], which is the only second-generation antidepressant registered for the treatment of depression in young people in the Netherlands and many other countries. Antidepressants may also be prescribed for anxiety, particularly in younger children, but no randomized placebo-controlled trial of citalopram for that purpose appears to have been conducted in children and adolescents, although fluoxetine has been found effective [355, 356].

Among the SSRIs, citalopram has also been most strongly associated with QT interval prolongation (particularly at higher doses), which may increase the risk for torsade de pointes and sudden cardiac death [357, 358] and which may be an additional safety-related reason, apart from treatment-emergent suicidality, to prefer fluoxetine as a first-line treatment.

The starting dose of antidepressants was generally higher than recommended. In particular, teens were usually prescribed an adult starting dose and were only rarely prescribed according to guidelines. Young children were prescribed according to the guidelines much more frequently (58%), but 10% of children were actually prescribed the adult starting dose, which is two to four times higher than the recommended dose. Few differences between prescribers were apparent, although specialists prescribed some SSRIs in slightly lower starting doses than GPs. This may be due to the slightly lower mean age of children receiving SSRIs from specialists compared to GPs.

Sertraline and citalopram were more likely to be prescribed according to the guidelines than other antidepressants. For sertraline, this is likely because the recommended starting dose is higher than that of other antidepressants, especially for older children. If we had used the stricter guideline rather than the more lenient guideline, adherence would have been markedly lower (23% overall).

For citalopram, the higher adherence to guidelines may be due to the availability of a liquid solution for citalopram, which facilitates low starting doses. In contrast, for fluoxetine, the tablet with the lowest dose currently available in the Netherlands contains 20 mg (1 DDD), which makes it difficult to provide the recommended dose of 5 mg. Although liquid fluoxetine was previously available, it is not currently on the Dutch market. The difficulty of providing low doses of fluoxetine may be one reason for physicians' preference for citalopram.

Several positive findings were also apparent. While GPs prescribed the majority of antidepressants in 1994 – 2003, prescriptions shifted to specialists over time, as recommended by guidelines. We also found that the starting doses of some antidepressants, particularly citalopram, decreased over the study period, suggesting increasing awareness among physicians of the importance of low starting doses in young people. This finding agrees with a previous study in the US showing increased prescription of low doses after the FDA warning in 2004 [351].

Finally, maintenance doses were nearly always in agreement with the guidelines; where they were not, this was usually because the dose was lower than recommended. In general, maintenance doses were very similar to starting doses. Up-titration from a low starting dose is recommended in the guidelines, but titration occurred in a minority of cases, probably because the starting dose was already within the maintenance range. Up-titration was more likely for second-generation antidepressants like citalopram and venlafaxine, for which a relatively low starting dose was also more likely.

The number of young people initiating antidepressant treatment decreased in the early 2000s, followed by a return to the level of 2001. Such a trend was also found in countries like the UK [346], but only to a slight extent or not at all in other countries, such as Canada [359] or Denmark [360]. The decrease in antidepressant initiation in young people was likely related to media coverage of the potential for treatment-emergent suicidality with antidepressant treatment [348], but this effect appears to have been transient.

Improving guideline adherence

Adherence to guidelines is often poor [361], and physicians' prescription choices are influenced by a multitude of other factors besides guidelines and continuing medical education. These influences may include the mass media (which may have been especially important with regard to the reduction in prescriptions for paroxetine) [348] and promotion by pharmaceutical companies [362]. A large body of research has examined barriers and facilitators to the implementation of guidelines in clinical practice [363, 364, 365]. Adherence is more likely when recommendations are specific and concrete rather than vague, when few additional resources are required for implementation, and when the evidence is strong and straightforward [365, 366].

While the recommendation to initiate antidepressant treatment in children with fluoxetine is highly specific and does not require any additional resources, the evidence base for the use of fluoxetine in young people is relatively limited, although stronger than that for other antidepressants [339], which may affect physicians' confidence in the recommendation.

Dedicated effort, for example implementation interventions [367], may be needed to improve adherence to guidelines. A variety of interventions have been found to increase guideline adherence, including provision of educational materials, audit and feedback, and reminders, but effects are modest [368]. Educational meetings, which are a common form of continuing medical education, also have small effects on improving guideline adherence [369]. A better understanding of the reasons behind physicians' preference for citalopram may help clarify how guideline adherence could be improved.

Strengths and limitations

This study has several strengths. First, use of a general population prescription database excludes the possibility of recall bias and selection bias. Another important strength is that we specifically examined first prescriptions, in contrast to many previous studies. Furthermore, we included a long time period of 21 years, which allowed us to examine time trends and the possible influence of major events, such as the recognition of a link between antidepressants and suicidality in young people in 2003 – 2004. This long time period also included very recent data (up to and including 2014).

Some limitations must also be acknowledged. An important limitation is that we did not have information about the indication for a prescription. As the guideline recommending fluoxetine is a guideline for the treatment of depression in young people, it may not apply to all prescriptions included herein.

In particular, amitriptyline was frequently prescribed in children and adolescents (approximately 15% of all prescriptions), even though tricyclic antidepressants are not recommended for the treatment of depression. Although we attempted to remove prescriptions for bed-wetting and pain, the remaining patients may still have been treated with amitriptyline for complaints other than depression.

A study among Dutch GPs suggested that SSRIs were usually prescribed for depression or anxiety, but tricyclics were often prescribed for bed-wetting, hyperactivity, tension headache or non-specific disease, and only rarely for depression [370]. Consequently, without information on the indication for amitriptyline prescriptions, it is difficult to determine whether these prescriptions were appropriate (although bed-wetting is the only approved indication for children and adolescents in the Netherlands).

However, as the majority of SSRI prescriptions to children and adolescents are for the purpose of treating depression [370], this limitation does not invalidate our finding that citalopram is preferred over fluoxetine, in contrast to the guideline. Furthermore, low

starting doses are important regardless of the indication and might even be of greater importance if antidepressants are prescribed for the treatment of anxiety, the most probable alternative indication for SSRIs, given the potential for increased anxiety early in treatment [371].

A second limitation of our study is that inpatient prescriptions are not included in the database. Consequently, some 'first prescriptions' may actually have been repeat prescriptions after treatment initiation during hospitalization. However, only 3 - 4% of all children who are treated in specialist mental health care are hospitalized in a year [372].

Conclusions

The guidelines on the treatment of depression in youth recommend fluoxetine as the treatment of choice. However, Dutch physicians appear to prefer citalopram over fluoxetine, even though citalopram has not been studied extensively and meta-analysis does not support its superiority over placebo in a pediatric population [339]. This is in contrast to findings from other countries, such as the United Kingdom, where antidepressant treatment in young people was most commonly initiated with fluoxetine (although citalopram has gained in popularity) [346]. Given that UK guidelines are similar to Dutch guidelines, this suggests that factors other than guidelines are likely to be the strongest driving forces behind (changes in) prescription patterns.

Furthermore, physicians tend to prescribe adult starting doses to older children. Although teens may weigh as much as adults, the possibility of a dose-response relationship with suicidality [341, 342] suggests that caution should be exercised, even for older children. The same may also apply to young adults, for whom antidepressants have also been shown to increase the risk of suicidality [191]. Although starting doses were adjusted for children and preteens, they were still frequently higher than recommended. Maintenance doses, on the other hand, were usually within the recommended range.

Taken together, these findings show that adherence to guidelines for antidepressant initiation in children and adolescents is poor. In light of the limited evidence for the efficacy of some antidepressants and the potential for treatment-emergent suicidality, physicians should be made aware of the importance of guideline adherence and cautious dosing of antidepressants in children and adolescents.

Appendix

Table 8.3: *Number (%) of prescriptions for each antidepressant*

		Incident users		
Antidepressant	Class	1994 – 2003	2004 – 2009	2010 - 2014
Amitriptyline	TCA	167 (14.0)	139 (17.1)	157 (16.8)
Clomipramine	TCA	52 (4.4)	8 (1.0)	4 (0.4)
Desipramine	TCA	9 (0.8)	1 (0.1)	0 (0.0)
Dosulepin	TCA	1 (0.1)	0 (0.0)	0 (0.0)
Doxepin	TCA	5 (0.4)	0 (0.0)	0 (0.0)
Imipramine	TCA	156 (13.1)	19 (2.3)	1 (0.1)
Maprotiline	TCA	3 (0.3)	0 (0.0)	1 (0.1)
Nortriptyline	TCA	5 (0.4)	5 (0.6)	8 (0.9)
Citalopram	SSRI	174 (14.6)	261 (32.0)	372 (39.9)
Escitalopram	SSRI	0 (0.0)	11 (1.3)	28 (3.0)
Fluoxetine	SSRI	121 (10.1)	123 (15.1)	184 (19.7)
Fluvoxamine	SSRI	116 (9.7)	31 (3.8)	30 (3.2)
Paroxetine	SSRI	266 (22.3)	95 (11.7)	19 (2.0)
Sertraline	SSRI	29 (2.4)	35 (4.3)	41 (4.4)
Agomelatine	Others	0 (0.0)	0 (0.0)	5 (0.5)
Bupropion	Others	3 (0.3)	11 (1.3)	13 (1.4)
Duloxetine	Others	0 (0.0)	3 (0.4)	7 (0.8)
Mianserin	Others	1 (0.1)	0 (0.0)	0 (0.0)
Mirtazapine	Others	26 (2.2)	27 (3.3)	36 (3.9)
Nefazodone	Others	4 (0.3)	0 (0.0)	0 (0.0)
Oxitriptan	Others	1 (0.1)	0 (0.0)	0 (0.0)
St. John's wort	Others	25 (2.1)	9 (1.1)	7 (0.8)
Trazodone	Others	2 (0.2)	7 (0.9)	5 (0.5)
Venlafaxine	Others	28 (2.3)	30 (3.7)	15 (1.6)

TCA: tricyclic antidepressants (ATC=N06AA); SSRI: selective serotonin reuptake inhibitor (ATC=N06AB); Others (ATC=N06AX)

Table 8.4: *Starting and maintenance doses of all antidepressants*

		Median dose in DDD/day (IQR)	
Antidepressant	Class	Starting	Maintenance
Amitriptyline	TCA	0.13 (0.13 – 0.33)	0.20 (0.13 – 0.33)
Clomipramine	TCA	0.20 (0.10 – 0.25)	0.25 (0.20 – 0.50)
Desipramine	TCA	0.50 (0.38 – 0.75)	0.75 (0.75 – 0.75)
Dosulepin	TCA	0.17	-
Doxepin	TCA	0.25 (0.18 – 0.33)	0.50
Imipramine	TCA	0.25 (0.20 – 0.25)	0.25 (0.25 – 0.50)
Maprotiline	TCA	0.25 (0.25 – 0.25)	-
Nortriptyline	TCA	0.13 (0.13 – 0.67)	0.67 (0.40 – 0.83)
Citalopram	SSRI	0.50 (0.30 – 1.00)	1.00 (0.50 – 1.00)
Escitalopram	SSRI	1.00 (0.50 – 1.00)	1.00 (1.00 – 1.00)
Fluoxetine	SSRI	1.00 (0.50 - 1.00)	1.00 (1.00 – 1.00)
Fluvoxamine	SSRI	0.50 (0.50 – 1.00)	1.00 (0.50 – 1.00)
Paroxetine	SSRI	1.00 (1.00 – 1.00)	1.00 (1.00 – 1.00)
Sertraline	SSRI	1.00 (1.00 – 1.00)	1.00 (1.00 – 2.00)
Agomelatine	Others	1.00 (1.00 – 1.50)	1.00 (1.00 – 1.00)
Bupropion	Others	0.50 (0.50 – 0.50)	0.50 (0.50 – 0.50)
Duloxetine	Others	0.50 (0.50 – 0.75)	0.75 (0.50 – 1.00)
Mianserin	Others	0.50	-
Mirtazapine	Others	0.50 (0.50 – 1.00)	0.50 (0.50 – 1.00)
Nefazodone	Others	0.50 (0.50 – 0.63)	-
Oxitriptan	Others	-	-
St. John's wort	Others	-	-
Trazodone	Others	0.33 (0.33 – 0.33)	0.33 (0.33 – 0.33)
Venlafaxine	Others	0.72 (0.375 – 0.75)	0.75 (0.73 – 0.75)
Overall	Others	0.50 (0.25 – 1.00)	1.00 (0.50 – 1.00)

No defined daily dose (DDD) has been set for St John's wort. TCA: tricyclic antidepressants (ATC=N06AA); SSRI: selective serotonin reuptake inhibitor (ATC=N06AB); Others (ATC=N06AX)

Part II

Who benefits from antidepressants?

Chapter 9

Influence of baseline severity on antidepressant efficacy for anxiety disorders: meta-analysis and meta-regression

Ymkje Anna de Vries, Peter de Jonge, Edwin van den Heuvel,
Erick H. Turner, Annelieke M. Roest

British Journal of Psychiatry (2016), 208, 515 - 521

Abstract

Background: Antidepressants are established first-line treatments for anxiety disorders, but it is not clear whether they are equally effective across the severity range.

Aims: To examine the influence of baseline severity of anxiety on antidepressant efficacy for generalized anxiety disorder (GAD), social anxiety disorder (SAD), obsessive-compulsive disorder (OCD), post-traumatic stress disorder (PTSD) and panic disorder (PD).

Methods: Fifty-six trials of second-generation antidepressants for the short-term treatment for an anxiety disorder were included. Baseline and change scores were extracted for placebo and treatment groups in each trial. Mixed-effects meta-regression was used to investigate the effects of treatment group, baseline severity, and their interaction.

Results: Increasing baseline severity did not predict greater improvement in drug groups compared to placebo groups. Standardized regression coefficients of the interaction term between baseline severity and treatment group were 0.04 (95% confidence interval -0.13 to 0.20, p=0.65) for GAD, -0.06 (-0.20 to 0.09, p=0.43) for SAD, 0.04 (-0.07 to 0.16, p=0.46) for OCD, 0.16 (-0.22 to 0.53, p=0.37) for PTSD, and 0.002 (-0.10 to 0.10, p=0.96) for PD. For OCD, baseline severity did predict improvement in both placebo and drug groups equally ($\beta = 0.11$, 95% confidence interval 0.05 to 0.17, p=0.001).

Conclusions: No relationship between baseline severity and the drug-placebo difference was found for anxiety disorders. These results suggest that if the efficacy of antidepressants is considered clinically relevant, they may be prescribed to anxious patients regardless of symptom severity.

Introduction

Anxiety disorders are the most common mental disorders, with a combined 12-month prevalence of 18.1% [4] and lifetime prevalence of 28.8% in the US [1]. Due to their high prevalence, combined with an often early onset and chronic course [1, 3, 373], anxiety disorders are the second most important cause of disability worldwide within the group of mental and behavioral disorders [316].

Antidepressants, including the second-generation selective serotonin reuptake inhibitors (SSRIs) and serotonin-norepinephrine reuptake inhibitors (SNRIs), have been found to be efficacious in the treatment of most anxiety disorders, including generalized anxiety disorder (GAD) [116], social anxiety disorder (SAD) [374], obsessive-compulsive disorder (OCD) [123], post-traumatic stress disorder (PTSD) [122], and panic disorder (PD) [119].

However, research in major depressive disorder (MDD) has suggested that the efficacy of antidepressants depends upon the baseline severity of depression. One meta-analysis of the Food and Drug Administration (FDA) database of randomized controlled trials (RCTs) of second-generation antidepressants showed that trials with higher mean baseline severity scores were more likely to be positive [70], while another found a significant interaction between baseline severity and treatment group in predicting improvement, such that the drug-placebo difference is relatively small at low levels of initial severity [69]. Similarly, an additional meta-analysis found no statistically significant difference between antidepressants and placebo in patients with subthreshold (minor) depression [375]. Among 4 analyses using individual patient data, 3 found that baseline severity of depression is associated with antidepressant efficacy [71, 72, 376], while another recent, large study did not find a significant association [76].

Much less is known about the relationship between baseline severity and antidepressant efficacy in the context of anxiety disorders. For OCD, a meta-analysis of 24 antidepressant RCTs found that baseline severity predicted greater improvement in both placebo and drug groups, but there was no evidence for an interaction between baseline severity and treatment group [78]. Recently, a meta-analysis of 12 RCTs of paroxetine for GAD and PD found no evidence for an interaction effect [77]. However, pooling trials for different disorders may obscure differences between disorders; it also necessitated the use of a secondary outcome for PD, rather than the primary, panic-specific outcome in these trials. Other single trials and small pooled analyses that have examined this question have reached contradictory conclusions [377, 378, 379].

The evidence to date, therefore, is (perhaps with the exception of OCD) conflicting, minimal, or absent altogether. To our knowledge, no study has comprehensively investigated whether baseline severity predicts antidepressant efficacy in all anxiety disorders. If antidepressant efficacy does depend upon baseline anxiety severity, this has important consequences for the continued development of guidelines for the treatment of anxiety disorders. Therefore, in order to investigate this question, we conducted a meta-analysis

and meta-regression of RCTs of SSRIs and SNRIs for the short-term treatment of anxiety disorders submitted to the Food and Drug Administration (FDA).

Methods

Study retrieval and data extraction

We obtained drug approval packages (a.k.a. reviews) for all SSRIs and SNRIs approved by the FDA for the short-term treatment of five anxiety disorders (GAD, SAD, OCD, PTSD and PD) [20]. Reviews were downloaded from the FDA website, when available, or requested from the FDA's Division of Freedom of Information [106].

From the FDA reviews, we extracted data on the duration of the trial, drug dose, number of participants, mean score on the primary outcome measure at baseline and endpoint (with standard error (SE) or standard deviation (SD) if available), and the mean change in the primary outcome measure (with SE/SD if available), for drug and placebo groups separately.

For GAD, the primary outcome measure in all included trials was the change from baseline on the HAM-A; for SAD, it was the change on the LSAS; for OCD, the change on the Yale-Brown Obsessive Compulsive Scale (Y-BOCS); and for PTSD, the change on the Clinician-Administered PTSD Scale, part 2 (CAPS-2).

For PD, the primary outcome for most trials was the number of panic attacks over a 1 – 3 week time frame, dichotomized into 0 (remitted) or ≥ 1 (not remitted); 4 trials, however, used the change in number of panic attacks as a primary outcome measure. Therefore, we also extracted data on the proportion of participants that were free of panic attacks at endpoint (remission rate). Furthermore, baseline scores were reported as the number of panic attacks in the 1, 2 or 3 weeks before baseline; we converted all scores to a 2-week time frame.

Data were extracted from the (modified) intention-to-treat analyses only, which used the last observation carried forward (LOCF) method to handle missing data from dropouts. Data were extracted preferably from the statistical review; however, we gave preference to other documents within the drug approval package (e.g. the medical review or administrative correspondence) if they provided more complete data, e.g. the SE/SD, as well as the mean change. If complete data was not available in the FDA review, we attempted to obtain the missing data from secondary sources (trial registries and published journal articles).

Outliers and missing data

Exploration of mean baseline scores, stratified by disorder, revealed no outliers in the dataset for GAD, SAD, OCD and PTSD. For PD, however, 1 placebo group and 1 drug group (from 2 separate trials) had mean baseline scores more than 2 standard deviations greater than the overall mean baseline score (after log-transformation to normalize the distribution); we therefore excluded these groups (but not other groups within the same trial) from our analysis.

We were able to obtain data on the mean change score and its SE/SD (or the remission rate for PD trials) for 46 out of 56 trials. For 40 trials, the FDA review provided complete data; for 2 trials, we obtained data on the SE/SD from the GlaxoSmithKline trial registry; and for 4 trials, we obtained data on the SE/SD (3 trials) or the remission rate (1 trial) from the matching published journal articles, after verifying that mean baseline and change scores matched those in the corresponding FDA review.

For 10 out of 56 trials, not all required data could be obtained from any source. Information on the SE/SD of the change score was missing for 1 PTSD and 6 OCD trials, while information on the remission rate was missing for 3 PD trials. For all 10 trials, however, information on the mean change score was available. For PTSD and OCD, the change score itself was strongly correlated with its SD in groups without missing data; we therefore imputed these missing SDs based upon the change score, group membership and their interaction. For PD, endpoint score was strongly correlated with the remission rate and we therefore imputed missing remission rates based upon the endpoint score, group membership and their interaction. Imputation was performed separately per disorder in SPSS 20, using multiple imputation with Fully Conditional Specification based upon linear regression in order to create 10 imputed datasets.

Statistical analysis

For our main analysis, we calculated effect sizes separately per treatment group, as a single-group pre-post effect size. This approach allowed us to investigate not only the relationship between baseline severity and antidepressant efficacy (drug-placebo difference), but also its underlying cause (i.e. change in the placebo response versus change in the drug response).

For GAD, SAD, OCD and PTSD, the standardized mean difference (SMD) was first calculated based on the (within-group) change score and its standard deviation, according to the formula $SMD = D/SD_D$, where D signifies the change score and SD_D the standard deviation of the change score (25).

By using this formula, we assume that the correlation between baseline and endpoint scores is 0.5, as the true correlation is unknown. We then applied Hedges' correction for

small sample size, where n indicates the number of participants in a group [380]:

$$g = (1 - \frac{3}{4(n-1)-1}) \times SMD$$

The standard error of Hedges' g was computed as follows [380]:

$$SE = (\frac{1}{n})(\frac{n-1}{n-3})(1 + n \times g^2) - (g^2/(1 - \frac{3}{4(n-1)-1})^2$$

For PD, we selected the remission rate itself as our effect measure.

In order to obtain a single effect size for the antidepressant arms of fixed-dose studies, we used a fixed-effects inverse-variance-weighted model to pool these drug groups into one estimate of effect size with its standard error for GAD, SAD, OCD and PTSD. A pooled remission rate was derived for PD by calculating the sample-size-weighted average of the remission rates in the different dose groups. For all disorders, a pooled baseline score was derived by calculating the sample-size-weighted mean of the baseline scores in the different dose groups. Pooling different dose groups may not be appropriate in the presence of a dose-response relationship, but with the exception of venlafaxine, such relationships usually cannot be demonstrated with second-generation antidepressants [381].

All analyses were performed in Stata 13. We performed meta-analyses using the *metan* command, applying a random-effects (DerSimonian-Laird) model to obtain summary statistics by disorder and group. In order to measure heterogeneity, the *heterogi* module within Stata was used to calculate I^2 and its 95% confidence interval [382].

Meta-regression was then performed separately per disorder using the *metareg* command. All meta-regressions were based on a mixed-effects model, used restricted maximum likelihood (REML) estimation of the residual between-study variance, and included group, baseline severity, and their interaction as predictors. The dependent variable was Hedges' g for GAD, SAD, OCD, and PTSD, and the remission rate for PD. Studies were weighted according to the inverse of their variance. For OCD, PTSD, and PD, meta-regression estimates from the 10 multiply imputed datasets were combined using the *mi* suite of commands in Stata.

Sensitivity analyses

As a secondary analysis, we calculated Hedges' g for the drug-placebo difference directly from the exact p-value for the statistical test performed at endpoint (or alternative methods as required [see reference 20]). The trial baseline severity score was calculated as the sample-size-weighted average of all groups (drug as well as placebo) included in the trial. We performed meta-regressions separately per disorder using the *metareg* com-

mand in Stata. The dependent variable was the drug-placebo difference (Hedges' g) for all disorders, and baseline severity was the only predictor in this analysis.

Additionally, to increase statistical power and improve generalizability of results, we expanded our dataset by including extra trials. We included active comparator arms and trials that were not conducted for the purpose of marketing approval, such as trials of other medications (e.g. antipsychotics) in which the antidepressant was used as an active comparator. Trials were obtained from the most recent meta-analyses examining (pharmacological) treatment of anxiety disorders [123, 187, 383]. As this introduced trials with very small sample sizes, additional heterogeneity, and likely reporting bias, this expanded set of trials was examined as a secondary analysis only.

We included parallel-group, placebo-controlled trials with a similar duration (8 – 16 weeks) as our primary set of trials. Trials with a sample that partly or fully overlapped with a trial included in the primary set were excluded, as were trials that did not use a compatible outcome (i.e., HAM-A, LSAS, Y-BOCS, CAPS, or number of panic attacks as a primary or secondary outcome). We repeated our primary analysis for this expanded set of trials.

Results

Description of included studies

A total of 21 reviews were obtained, including nine (formulations of) SSRIs and SNRIs: escitalopram, duloxetine, fluoxetine, fluvoxamine, fluvoxamine controlled release (CR), paroxetine, paroxetine CR, sertraline and venlafaxine extended release (XR). These reviews comprised a total of 59 RCTs: 11 for GAD, 11 for SAD, 13 for OCD, 7 for PTSD and 17 for PD. However, we excluded 2 trials for SAD and 1 trial for PD, as baseline severity scores were incomparable to the other trials. Specifically, the excluded SAD trials did not use the Liebowitz Social Anxiety Scale (LSAS), while the excluded PD trial failed to distinguish between full and limited-symptom panic attacks.

Consequently, 56 trials were included in this study. We only used data from doses recommended by the FDA; therefore, for paroxetine, we excluded the 20-mg dose in 1 OCD trial and the 10- and 20-mg dose in 1 PD trial, as only doses of 40 mg and higher were judged to be effective for these disorders.

All included trials were short-term, randomized, double-blind and placebo-controlled. Some trials also utilized active comparators, but data from these were not included in our primary analyses. Trial duration was 12 weeks for all SAD and PTSD trials, 8 – 10 weeks for GAD trials, 8 – 16 weeks for OCD trials, and 10 – 12 weeks for PD trials.

Trials included adults of 18 years or older (with the exception of 5 OCD trials that included adolescents together with adults) who met DSM-III-R or DSM-IV criteria for

the anxiety disorder under investigation. The majority of trials (71%) used a flexible-dose design, allowing investigators to titrate the dose according to a subject's response, while the remainder (29%) used a fixed-dose design with patients randomized to one of 2 or 3 different dosage groups.

The 56 trials included a total of 55 placebo groups and 78 drug groups, which were pooled into 56 drug groups. The total number of participants across disorders was 14,710, of whom 6,386 were randomized to placebo and 8,324 were randomized to drug. Baseline severity was generally in the moderate to severe range. Participant numbers and baseline severity scores by disorder and treatment group are shown in Table 9.1. A complete listing of included trials, with baseline and change scores (for GAD, SAD, OCD, and PTSD) or remission rates (for PD), is provided in Tables 9.4 and 9.5.

Table 9.1: *Participant numbers and baseline scores by disorder and group.*

	Participants			Mean baseline score (range)	
Disorder	Placebo	Drug	Total	Placebo	Drug
GAD	1628	2219	3847	23.9 (22.1 – 25.9)	24.0 (22.5 – 26.0)
SAD	1255	1379	2634	85.8 (73.3 – 93.9)	86.7 (78.0 – 95.9)
OCD	976	1583	2559	24.5 (22.6 – 26.3)	24.4 (22.6 – 26.6)
PTSD	829	981	1810	74.3 (72.0 – 78.4)	74.4 (72.0 – 77.4)
PD	1698	2162	3860	10.6 (6.2 – 19.2)	11.5 (6.9 – 17.6)

GAD: generalized anxiety disorder; OCD: obsessive-compulsive disorder; SAD: social anxiety disorder; PD: panic disorder; PTSD: post-traumatic stress disorder. Baseline scores for GAD are based on the Hamilton Anxiety Rating Scale (HAM-A), for SAD on the Liebowitz Social Anxiety Scale (LSAS), for OCD on the Yale-Brown Obsessive-Compulsive Scale (Y-BOCS), for PTSD on the Clinician-Administered PTSD Scale (CAPS-2) and for PD on the number of panic attacks in the 2 weeks before baseline.

Meta-analysis

We performed a meta-analysis of effect sizes, stratified by disorder and group (Table 9.2). The placebo group effect size was smallest for OCD (0.49), followed by SAD (0.65). These placebo effects were 59% and 60%, respectively, of the effects found for the corresponding drug groups. By contrast, for GAD and PTSD, the placebo effect sizes were 1.03 and 0.97, respectively; these effects were 76% and 83%, respectively, of the effects found for the corresponding drug groups. For PD, the placebo remission rate was 45%, which was 76% of the rate found in the drug group. Substantial heterogeneity was present for nearly all groups, with I^2 ranging from 46% for the placebo groups in SAD trials to 89% for the drug groups in PTSD trials, although confidence intervals were generally wide.

Table 9.2: *Meta-analysis of effect sizes*

	Effect size (95% CI)		
Disorder	Placebo	Drug	Difference
GAD	1.03 (0.93 - 1.13)	1.35 (1.24 - 1.46)	0.32 (0.16 - 0.47)
SAD	0.65 (0.57 - 0.74)	1.08 (0.99 - 1.18)	0.43 (0.29 - 0.57)
OCD	0.49 (0.39 - 0.59)	0.83 (0.73 - 0.92)	0.34 (0.19 - 0.48)
PTSD	0.97 (0.81 - 1.13)	1.17 (0.91 - 1.44)	0.20 (-0.15 - 0.55)
PD	0.45 (0.38 - 0.52)	0.59 (0.55 - 0.65)	0.14 (0.06 - 0.22)

GAD: generalized anxiety disorder; OCD: obsessive-compulsive disorder; SAD: social anxiety disorder; PD: panic disorder; PTSD: post-traumatic stress disorder. The effect size is expressed as Hedges' g for GAD, SAD, OCD and PTSD, and as remission rate for PD.

Meta-regression

For both GAD and SAD, neither baseline severity nor the interaction between group and baseline severity were statistically significant predictors of the effect size, although group membership was (p=0.001 and <0.001, respectively) (Table 9.3 and Figure 9.1). For OCD, the interaction between baseline severity and group membership was not significant, but the main effects of both group membership and baseline severity were (group: p<0.001; baseline: p=0.001), indicating that the (positive) slope of the association between baseline severity and effect size was similar in the placebo and drug groups. For PTSD, none of the predictors achieved statistical significance.

For PD, we modeled the relationship between baseline severity and the remission rate. In this model, the interaction between group membership and baseline severity was not significant (Table 9.3 and Figure 9.2). Group membership was a significant predictor of the remission rate (p=0.007), while baseline severity was not.

Paralleling these results, we found that including treatment group in the model reduced between-group heterogeneity for GAD (I^2 decreasing from 83% to 67%), SAD (84% to 49%), OCD (75% to 52%), and PD (86% to 77%), while it did not reduce heterogeneity for PTSD (85% to 84%). Including the main effect of baseline reduced heterogeneity for OCD only (52% to 14%), while including the interaction did not reduce heterogeneity further for any disorder.

Sensitivity analyses

In our secondary analysis, baseline severity was not a significant predictor of the drug-placebo difference for any disorder, and the effect of baseline severity was positive (i.e. in the expected direction) only for PTSD ($\beta = 0.084$, p=0.42). For the other disorders, the effect ranged from -0.064 to -0.002 (p-values ranging from 0.170 to 0.860).

Table 9.3: *Meta-regression analysis*

		Model 1 (with interaction)		Model 2 (without interaction)	
Disorder	Predictor	β (95% CI)	p	β (95% CI)	p
GAD	Group	0.31 (0.15 - 0.47)	0.001	0.32 (0.16 - 0.48)	0.001
	Baseline	-0.03 (-0.15 - 0.09)	0.60	-0.01 (-0.09 - 0.07)	0.77
	G x B	0.04 (-0.13 - 0.20)	0.65		
SAD	Group	0.43 (0.29 - 0.57)	0.001	0.43 (0.29 - 0.56)	0.001
	Baseline	0.06 (-0.04 - 0.17)	0.21	0.04 (-0.03 - 0.11)	0.29
	G x B	-0.06 (-0.20 - 0.09)	0.43		
OCD	Group	0.35 (0.0.23 - 0.46)	0.001	0.35 (0.24 - 0.47)	0.001
	Baseline	0.09 (-0.004 - 0.17)	0.04	0.11 (0.05 - 0.17)	0.001
	G x B	0.04 (-0.07 - 0.16)	0.46		
PTSD	Group	0.20 (-0.17 - 0.56)	0.25	0.20 (-0.16 - 0.56)	0.24
	Baseline	-0.01 (-0.28 - 0.27)	0.96	0.07 (-0.12 - 0.26)	0.41
	G x B	0.16 (-0.22 - 0.53)	0.37		
PD	Group	0.14 (0.05 - 0.24)	0.006	0.14 (0.05 - 0.24)	0.005
	Baseline	-0.01 (-0.11 - 0.09)	0.85	-0.01 (-0.08 - 0.06)	0.81
	G x B	0.002 (-0.10 - 0.10)	0.96		

GAD: generalized anxiety disorder; OCD: obsessive-compulsive disorder; SAD: social anxiety disorder; PD: panic disorder; PTSD: post-traumatic stress disorder. G x B = Group x Baseline interaction.

The expanded set of trials included 9 additional trials for GAD [384, 385, 386, 387, 388, 389, 390, 391, 392], 10 additional trials for SAD [393, 394, 395, 396, 397, 398, 399, 400, 401, 402], 6 additional trials for OCD [403, 404, 405, 406, 407, 408], 6 additional trials for PTSD [409, 410, 411, 412, 413, 414], and 6 additional trials for PD [415, 416, 417, 418, 419, 420]. The range of baseline severity was increased slightly to substantially for all disorders in this expanded set of trials. However, no statistically significant interaction effects were found for any disorder, although the main effect of baseline became statistically significant for GAD, PTSD, and SAD, in addition to OCD (Table 9.6 in the Appendix).

Discussion

Principal findings

We found no evidence that baseline severity of disorder affects the efficacy of second-generation antidepressants in the short-term treatment of anxiety disorders. This finding stands in remarkable contrast to what has been reported in studies investigating MDD.

For OCD, baseline severity did predict change in score in both placebo and drug groups, but no differential effect was apparent. This suggests that patients with more severe OCD may show substantially greater improvement with antidepressant treatment than

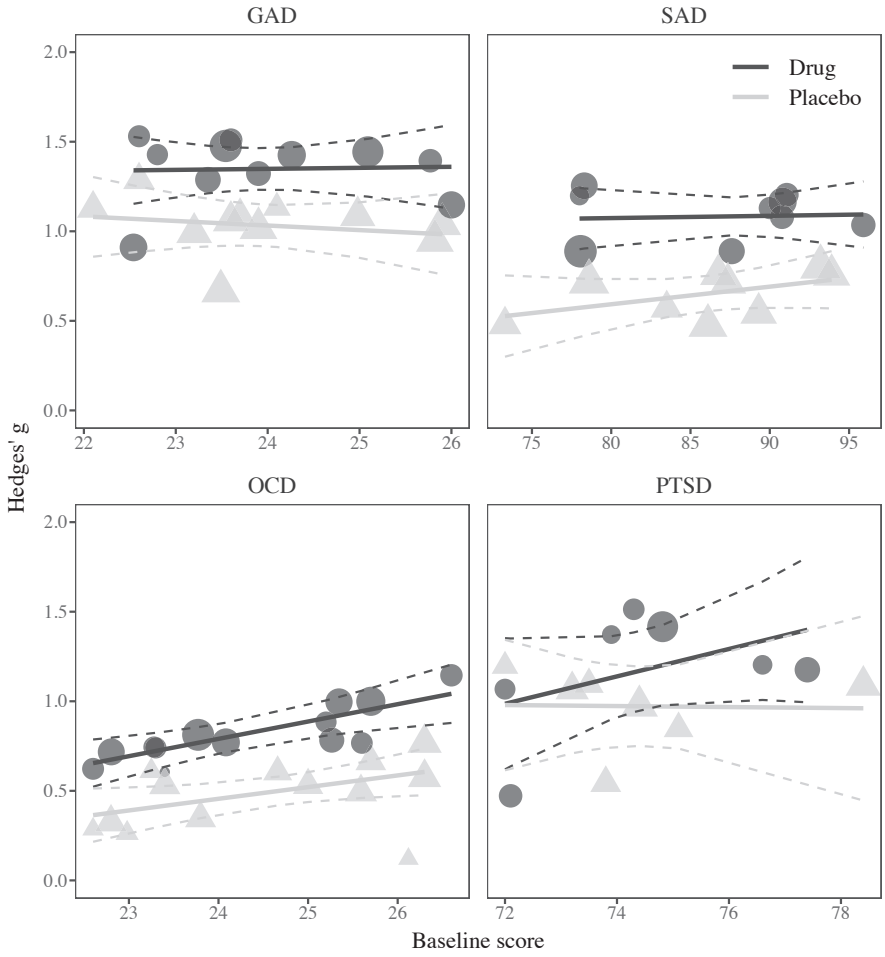

Figure 9.1: *Meta-regression analysis for GAD, SAD, OCD, and PTSD. Data points are sized proportionally to the inverse of their standard error.*

patients with milder OCD, but this is not due to improved antidepressant efficacy in severe OCD. Our sensitivity analysis, in which 37 additional trials were included, suggests that a similar regression to the mean effect might also occur in other anxiety disorders, but confirmed the lack of evidence for an interaction effect in all anxiety disorders.

Effect sizes for the drug-placebo difference were unaffected by inclusion of baseline severity as a main effect or in interaction with group, but they were generally smaller than a criterion for clinical significance previously used (though without clear justification) for MDD (d=0.5) [69]. However, it has been shown that effect sizes exceeding 0.5 are not achieved by most current treatments, either in psychiatry or in general medicine [421].

Furthermore, clinical significance is context- and disorder-specific [422]. An empirically-

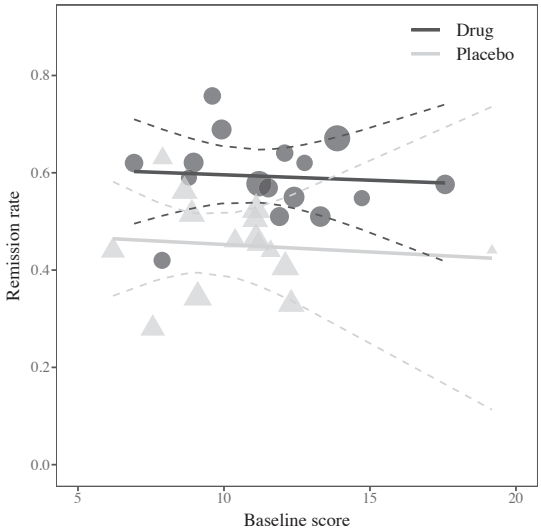

Figure 9.2: *Meta-regression analysis for PD. Data points are sized proportionally to the inverse of their standard error.*

derived criterion of d=0.24 has been suggested to be a more meaningful threshold for clinically significant efficacy for MDD [423]. Although not all included anxiety disorders met that threshold in our primary analysis, the drug-placebo difference may have been underestimated slightly due to our analytical approach. We have previously analyzed the trials included in this study by conventional meta-analytic methods and found effect sizes of 0.27 and greater [20]. Further research is required to establish whether antidepressant efficacy in the different anxiety disorders may be considered clinically significant in all or in subsets of patients.

Additionally, the expected efficacy of medication is only one of many factors that play a role in the decision to prescribe antidepressants for an individual patient. Other issues, such as the expected burden of side effects, the burden posed by the disorder itself, the anticipated course of the disorder, and the availability, acceptability, and efficacy of alternative treatment options like psychotherapy, must also be considered [424]. These considerations might lead to different prescribing decisions for patients with mild versus severe anxiety disorders, even in the absence of differential efficacy.

Comparison with depression

For MDD, it has been suggested that the threshold for "antidepressant-treatable depression" should be higher [425], given that mildly depressed patients do not show a strong differential response to antidepressants over placebo. It is possible that this threshold needs to be even higher in anxiety disorders, and that we would find greater efficacy of

antidepressants in the most severely anxious patients than we do in patients with less extreme anxiety.

Another possible explanation lies in the difference in chronicity between anxiety disorders and MDD. Depressive episodes tend to be relatively short, with a median duration of 3 – 6 months [3, 426], while the median duration of an anxiety episode has been estimated at 16 months [3]. However, duration of a depressive episode is positively correlated with severity [3]. This suggests that, all other things being equal, patients with mild depression are more likely to remit spontaneously within the short time frame of a clinical trial than patients with severe depression.

Since a drug effect cannot be demonstrated in patients who remit spontaneously, efficacy would be expected to be reduced in a patient group with a high likelihood of spontaneous remission. Consequently, the correlation between severity and episode duration may explain the increase in antidepressant efficacy with increasing depression severity. Although severity also correlates with episode duration in anxiety disorders [427], this may play less of a role within the context of a short clinical trial. Given the long median duration of an episode of anxiety, spontaneous remission rates would be expected to be relatively low across the severity range.

Alternatively, we may have been unable to detect a relationship between baseline severity and antidepressant efficacy due to a restriction in range. As trials generally have a minimum severity requirement, as well as exclusion criteria that tend to exclude severely ill patients (e.g., regarding treatments received and comorbidity), baseline severity was restricted to the moderate to very severe range in our primary analyses, although the range was extended somewhat in our sensitivity analyses. However, previous studies in MDD, which were able to detect a relationship between baseline severity and antidepressant efficacy, had similarly restricted severity ranges.

We may also have had insufficient power to find a statistically significant interaction: although the total number of trials included in the current study was large, the number of trials per disorder was limited to 7 – 16. However, the results do not suggest a meaningful trend toward increasing efficacy with increasing severity, except possibly for PTSD, for which we had the fewest available trials. The estimates of the interaction effect were substantially lower (≤ 0.04, except for PTSD) or even in the opposite direction as the estimate previously reported for depression [69]. These estimates are small enough to be of limited clinical relevance, although some confidence intervals encompass values that may be considered clinically relevant.

Furthermore, our secondary analyses, which included 6 to 10 additional trials per disorder, also showed no evidence for increasing efficacy with increasing severity, even for PTSD. Larger samples would be required to conclusively exclude the possibility of even small interaction effects, but unfortunately the number of randomized trials that have been conducted is limited, and consequently a larger sample of trials is not available.

Strengths and limitations

Among the strengths of the current study is the fact that we were able to obtain a dataset that is free from the influence of publication or outcome reporting bias. An additional strength of the study is the high quality of the included trials. As these trials were conducted for the purpose of obtaining marketing approval, they were required to meet strict standards on internal validity (blinding, randomization, etc.). This also ensured that trials conducted for the treatment of the same disorder were, in general, quite comparable.

A limitation is that we were forced to use a different outcome measure for PD (remission) than we did for the other disorders; consequently, we cannot rule out the possibility that an interaction might have been found if we had used a continuous outcome (such as change) instead.

An additional, important limitation is that we did not have access to individual patient data and hence used summary data instead. Use of summary data necessarily causes a loss of information, due to averaging out inter-individual variation within trials, and a concomitant loss of power. Although individual patient data have historically been very difficult to obtain, novel initiatives are now beginning to increase their accessibility to researchers (e.g. clinicalstudydatarequest.com).

Given our current results, which suggest that the influence of baseline severity on antidepressant efficacy may be different for anxiety disorders than for depression, future research making use of individual patient data is essential to provide a definitive answer to this important question.

Conclusions

In conclusion, we found no evidence for an interaction between treatment group and baseline severity in predicting symptom change. It has previously been recommended that treatment with antidepressants should be withheld for mild depression [425, 428]. Our results show that this cannot be simply extrapolated to anxiety disorders, and it would therefore be premature to recommend that antidepressants be withheld for mild anxiety. What defines a clinically relevant effect size remains a matter of debate, but if the effect of antidepressants on anxiety is considered clinically relevant, these results suggest that antidepressants may be prescribed to anxious patients regardless of symptom severity.

Appendix

Table 9.4: *Characteristics of included trials for GAD, SAD, OCD and PTSD*

			Placebo group			Antidepressant group		
Disorder	Drug	Trial	N	Baseline	Change (SD)	N	Baseline	Change (SD)
GAD	Escitalopram	MD-05 [130]	128	22.1	7.7 (6.8)	124	22.8	9.6 (6.7)
		MD-06 [130]	138	22.6	7.6 (5.9)	143	22.6	9.2 (6.0)
		MD-07 [131]	153	23.2	7.4 (7.4)	154	23.6	11.3 (7.4)
	Paroxetine	641 [132]	180	23.9	9.6 (9.4)	188	23.8	12.5 (8.2)
						197	23.3	12.2 (8.4)
		642 [133]	163	23.6	9.5 (8.9)	161	23.9	11.8 (8.9)
		637 (N/A)	183	25.9	11.3 (10.8)	181	26.0	12.4 (10.8)
	Duloxetine	HMBR [134]	173	25.8	8.4 (8.8)	165	25.1	12.8 (8.7)
						169	25.1	12.5 (8.7)
		HMDT [135]	158	23.5	5.9 (8.8)	161	22.5	8.1 (8.9)
		HMDU [136]	158	25.0	9.2 (8.4)	149	25.8	11.8 (8.4)
	Venlafaxine XR	210 [137]	96	24.1	9.5 (8.3)	86	24.7	11.1 (8.8)
						81	24.5	11.7 (7.8)
						86	23.6	12.1 (7.5)
		214 [138]	98	23.7	8 (7.2)	87	23.7	10.6 (7.6)
						87	23.0	9.8 (8.0)
SAD	Fluvoxamine	3107 [149]	125	89.3	13.2 (24.1)	110	90.0	26.6 (23.4)
		3108 [150]	148	93.9	26.2 (34.4)	126	95.9	34.6 (33.2)
	Paroxetine	502 [151]	145	86.1	15.6 (32.8)	136	87.6	29.4 (32.9)
		382 [152]	92	83.5	14.5 (25.2)	90	78	30.5 (25.2)
		454 [153]	92	73.3	15.0 (31.1)	89	79.8	31.4 (29.5)
						88	77.5	24.5 (30.3)
						91	76.9	25.2 (30.0)
	Paroxetine CR	790 [154]	184	78.6	17.6 (24.4)	185	78.3	31 (24.6)
	Sertraline	R-0601 [155]	196	93.2	21.4 (26.6)	205	90.8	31.3 (26.8)
	Venlafaxine XR	387 [158]	138	86.8	19.9 (26.1)	133	91.1	31 (25.6)
		393 [159]	135	87.4	22.1 (30.9)	126	90.8	32.8 (30.2)
OCD	Fluoxetine	HCEP 1 [165]	47	23.0	1.2 (4.5)	47	22.9	5.5 (7.1)
						45	22.4	4.3 (5.3)
						47	23.1	4.2 (6.7)
		HCEP 2 [165]	41	26.1	0.6 (4.6)	39	24.4	3.5 (5.9)
						41	25.4	6.9 (8.1)
						42	26.0	9.1 (9.4)
		E079 [166]	56	23.3	3.7 (6.0)	52	23.8	5.1 (6.4)
						52	25.5	4.7 (6.9)
						54	23.0	6.1 (6.9)
	Fluvoxamine	5529 (N/A)	80	22.8	1.7 (?)	79	23.3	4.9 (?)
		5534 [167]	77	23.8	1.7 (4.9)	78	22.6	4.0 (6.3)

continued

Table 9.4: *Characteristics of included trials for GAD, SAD, OCD and PTSD*

			Placebo group			Antidepressant group		
Disorder	Drug	Trial	N	Baseline	Change (SD)	N	Baseline	Change (SD)
	Fluvoxamine CR	3103 [168]	119	26.3	5.9 (7.6)	113	26.6	8.7 (7.5)
	Paroxetine	116 [169]	88	25.6	3.4 (6.8)	83	25.4	6.3 (6.7)
						83	25.3	7.3 (6.7)
		118 (N/A)	75	24.7	4.6 (7.5)	79	23.3	5.6 (7.5)
		136 [170]	99	26.3	3.9 (?)	198	25.7	6.9 (?)
	Sertraline	237/248 [171]	44	22.6	1.5 (?)	43	23.4	3.8 (?)
		371/372 [172]	84	23.4	3.4 (?)	79	23.2	6.0 (?)
						81	24.6	4.5 (?)
						80	23.5	6.2 (?)
		546 [173]	79	25	3.6 (?)	85	25.2	6.5 (?)
		495 (N/A)	87	25.7	5.0 (?)	83	25.6	5.4 (?)
PTSD	Sertraline	641 [160]	82	73.8	15.4 (28.1)	84	72.1	13.1 (27.5)
		682 (N/A)	94	72.0	27.9 (?)	94	72.0	27.4 (?)
		640 [161]	104	73.5	26.2 (23.8)	98	73.9	33 (23.9)
		671 [162]	90	75.1	23.2 (27.1)	93	76.6	33 (27.2)
	Paroxetine	651 [163]	167	74.4	25.3 (25.8)	166	75.3	39.6 (25.8)
						156	74.3	37.9 (28.7)
		648 [164]	133	73.2	24.7 (23.1)	136	74.3	35.5 (23.3)
		627 (N/A)	159	78.4	26.2 (24.0)	154	77.4	30.8 (26.1)

CR: controlled release; GAD: generalized anxiety disorder; OCD: obsessive-compulsive disorder; SAD: social anxiety disorder; PTSD: post-traumatic stress disorder; XR: extended release.

Table 9.5: *Characteristics of included trials for panic disorder*

		Placebo group			Antidepressant group		
Drug	Trial	N	Baseline	Remission rate	N	Baseline	Remission rate
Fluoxetine	HCJC [429]	90	7.6	28	90	7.9	42
	HCJB (N/A)	104	6.2	44	107	6.9	62
Paroxetine	120 [139]	69	11.6	43.9	72	9.6	75.8
	187 [141]	123	12.3	33	123	11.9	51
	223 (N/A)	68	7.9	63	77	8.8	59
Paroxetine CR	494 [142]	129	11.1	50.4	122	9.9	68.9
	495 [142]	136	8.9	51.5	123	11.5	56.9
	497 [142]	130	8.7	56.2	132	9.0	62.1
Sertraline	629 [143]	87	10.4	46	79	12.8	62
	630 [144]	88	11.2	?	88	12.1	?
	529 [129]	-	-	-	42	20.3	?
					41	21.3	?
					44	11.5	?
	514 (N/A)	38	19.2	?	38	14.1	?
					36	15.4	?
Venlafaxine XR	398 [145]	156	9.1	34.4	158	11.0	54.1
					159	11.4	61.4
	399 [146]	157	11.1	46.5	156	15.7	64.1
					160	12.1	70.0
	353 [147]	155	12.1	40.6	155	13.3	51.0
	391 [148]	168	11.1	52.4	160	12.4	55.0

CR: controlled release; XR: extended release.

Table 9.6: *Secondary meta-regression analysis with expanded set of trials*

		Model 1 (with interaction)		Model 2 (without interaction)	
Disorder	Predictor	β (95% CI)	p	β (95% CI)	p
GAD	Group	0.35 (0.17, 0.52)	<0.001	0.34 (0.17, 0.51)	<0.001
	Baseline	0.11 (-0.03, 0.25)	0.13	0.13 (0.04, 0.23)	0.007
	G x B	0.05 (-0.13, 0.23)	0.57		
SAD	Group	0.47 (0.35, 0.59)	<0.001	0.45 (0.33, 0.58)	<0.001
	Baseline	0.15 (0.04, 0.25)	0.011	0.10 (0.02, 0.18)	0.012
	G x B	-0.08 (-0.23, 0.07)	0.29		
OCD	Group	0.35 (0.23, 0.47)	<0.001	0.36 (0.23, 0.48)	<0.001
	Baseline	0.14 (0.03, 0.25)	0.013	0.18 (0.10, 0.27)	<0.001
	G x B	0.09 (-0.08, 0.25)	0.28		
PTSD	Group	0.21 (-0.03, 0.45)	0.08	0.21 (-0.03, 0.44)	0.08
	Baseline	0.17 (-0.06, 0.40)	0.14	0.20 (0.05, 0.35)	0.013
	G x B	0.06 (-0.26, 0.37)	0.71		
PD	Group	0.13 (0.06, 0.20)	0.001	0.13 (0.06, 0.20)	0.001
	Baseline	0.00 (-0.07, 0.07)	0.97	-0.03 (-0.07, 0.02)	0.20
	G x B	-0.05 (-0.13, 0.04)	0.30		

GAD: generalized anxiety disorder; OCD: obsessive-compulsive disorder; SAD: social anxiety disorder; PD: panic disorder; PTSD: post-traumatic stress disorder. G x B = Group x Baseline interaction.

Chapter 10

Initial severity and antidepressant efficacy for anxiety disorders: an individual patient data meta-analysis

Ymkje Anna de Vries, Annelieke M. Roest,
J. G. M. (Hans) Burgerhof, Peter de Jonge

Submitted

Abstract

Objective: To examine the influence of initial severity on antidepressant efficacy for generalized anxiety disorder (GAD), social anxiety disorder (SAD), obsessive-compulsive disorder (OCD), post-traumatic stress disorder (PTSD), and panic disorder (PD).

Methods: Individual patient data of 8,979 participants in 29 antidepressant trials were requested from Clinical Study Data Request. Mixed-effects models were used to investigate an interaction between initial severity and treatment group.

Results: For GAD, an interaction between treatment group and severity was found. The antidepressant-placebo difference was 1.4 (95% CI: 0.4-2.5, SMD: 0.15) points on the Hamilton Anxiety Rating Scale (HAM-A) for mildly ill participants (baseline HAM-A of 10), increasing to 4.0 (95% CI: 3.4-4.6, SMD: 0.43) or greater for severely ill participants (baseline HAM-A of 30).

For SAD, OCD, and PTSD, no interaction was found. Across severity levels, the mean difference was 16.1 (95% CI: 12.9-19.3, SMD: 0.59) on the Liebowitz Social Anxiety Scale (LSAS) for SAD, 3.4 (95% CI: 2.5-4.4, SMD: 0.39) on the Yale-Brown Obsessive-Compulsive Scale (Y-BOCS) for OCD, and 10.3 (95% CI: 6.9-13.6, SMD: 0.41) on the Clinician-Administered PTSD Scale (CAPS) for PTSD.

For PD, the antidepressant-placebo difference in number of panic attacks/2 weeks was 0.4 (95% CI: 0.3-0.6) for participants with 10 panic attacks/2 weeks at baseline, increasing to 0.9 (95% CI: 0.7-1.2) for participants with 20, and to 4.7 (95% CI: 3.0-6.4) for participants with 40.

Conclusions: Antidepressants are equally effective across the severity range for SAD, OCD, and PTSD. For GAD and PD, however, antidepressant benefits are small at low severity.

Registration: NCT02476136

Introduction

Antidepressants are considered effective treatments for major depressive disorder (MDD) [11, 19] and anxiety disorders [12, 123, 187, 264]. However, research in MDD has suggested that antidepressant efficacy may depend upon initial symptom severity. Both trial-level [69, 70, 375] and individual patient data (IPD) meta-analyses [71, 72, 376] have found that antidepressants provide few benefits compared to placebo for patients with low initial severity. Consequently, many guidelines no longer recommend antidepressants for mild depression [74, 428].

A relationship between initial severity and efficacy has also been found for the use of antipsychotics in schizophrenia [430], which suggests that this is a cross-diagnostic phenomenon. Recently, however, two IPD meta-analyses, both substantially larger than previous analyses, did not find an association between initial severity and antidepressant efficacy for MDD [75, 76], indicating that this question is not yet settled.

Antidepressants are also commonly used for anxiety disorders [23], but comparatively little evidence is available for these disorders. Trial-level meta-analyses for OCD [78], generalized anxiety disorder (GAD) and panic disorder (PD) [77], and social anxiety disorder (SAD) [431, 432] found no evidence that antidepressant efficacy increases with increasing severity. We also found no support for an association between initial severity and treatment efficacy in a recent meta-analysis of 56 antidepressant trials for GAD, SAD, OCD, PTSD, and PD [433].

However, trial-level meta-regression analyses that examine a participant-level variable (such as initial severity) may be prone to the ecological fallacy [434], in which a trial-level relationship can be found that does not exist at the participant level, or vice versa. They can also be underpowered and suffer from a restriction of range, since using the mean baseline severity across participants will average out extreme scores. Hence, IPD is needed to provide better insight into whether initial severity is associated with antidepressant efficacy for anxiety disorders.

However, few studies have used IPD and most of these studies had significant limitations. Two studies examined efficacy in subgroups of less and more severely anxious patients (for GAD and SAD) without actually testing for differences between the subgroups [435, 436]. Two other patient-level analyses for GAD tested the association between severity and dichotomized outcomes, with one analysis reporting an association only between severity and remission [437] and the other only between severity and response [379]. Two patient-level analyses for SAD also found contradictory results, with one reporting greater efficacy in more severely anxious participants than in less severely anxious participants [378] while the other reported similar efficacy [438]. Finally, a post-hoc analysis of a trial for PTSD found no evidence for moderation by baseline severity, but this was a negative trial, which may have made it impossible to detect a moderation effect [439]. To our knowledge, there are no patient-level analyses for OCD or PD.

Given the limitations of the available evidence (including dichotomization of outcomes and predictors, which leads to a significant loss of power [440]) and the contradictory results, the question of whether initial severity moderates antidepressant efficacy for anxiety disorders, OCD, and PTSD remains unanswered. In the current study, we therefore examined this question using IPD from 29 trials enrolling 8,979 participants.

Methods

Data source

We requested IPD from Clinical Study Data Request, a multi-sponsor data-sharing platform [441]. We first identified all selective serotonin reuptake inhibitors (SSRIs) and serotonin-norepinephrine reuptake inhibitors (SNRIs) developed by participating sponsors. These included paroxetine, fluoxetine, and duloxetine. We then identified all double-blind RCTs of these antidepressants for an anxiety disorder that were mentioned in Food and Drug Administration drug approval packages [20] or the GlaxoSmithKline [442] and Lilly [443] trial registries. We included only RCTs that were placebo-controlled, short-term ($\leq$ 16 weeks), and performed primarily in adults.

Primary outcomes

As our primary outcome, we chose the outcome usually considered primary for that disorder. For GAD, this was the Hamilton Rating Scale for Anxiety (HAM-A); for SAD, the Liebowitz Social Anxiety Scale (LSAS); for OCD, the Yale-Brown Obsessive-Compulsive Scale (Y-BOCS); and for PTSD, the Clinician-Administered PTSD Scale (CAPS). For PD, most of the included trials used response as an outcome (defined as having 0 full panic attacks), but we selected the number of full panic attacks per 2 weeks, since dichotomizing a continuous outcome leads to a significant loss of information.

Patient population

We included patients with a valid baseline score on the primary outcome and at least one valid follow-up score. Patients assigned to placebo, the investigative antidepressant, or a comparator SSRI or SNRI were included. We excluded patients assigned to other active comparators (e.g. benzodiazepines).

Statistical analysis

We conducted separate analyses for each disorder. For GAD, SAD, OCD, and PTSD, we applied linear mixed models, using the nlme package (version 3.1-127) for R (version 3.3.0). The effect measure of interest was the change from baseline on the primary outcome. The initial model included all fixed effects, regardless of significance. These were initial severity, treatment group, linear and quadratic terms for time (in days since baseline), and their two- and three-way interactions. Baseline and change scores were grand-mean centered and standardized, while time was centered at trial endpoint and standardized.

Using this first model, we modeled the covariance structure of the nested data (observation within participant within trial). We considered a random intercept at the trial level and a random intercept plus random effects for linear and quadratic time at the participant level. For these random effects, we examined compound symmetry, diagonal, and unstructured covariance matrices. Additionally, we considered an autocorrelation term for the residuals. We used restricted maximum likelihood (REML) for estimation and the Akaike Information Criterion (AIC) to select the best-fitting covariance structure.

Subsequently, we refitted the model using maximum likelihood (ML) and removed the least significant fixed effects by backward selection. If an interaction or quadratic effect was significant, we retained all component main or linear effects regardless of significance. We used the AIC to select the best-fitting model. However, for clarity we further simplified models containing non-significant terms even if there was a marginal AIC difference in favor of the more complex model. In these cases, both the Bayesian Information Criterion (BIC) and (when applicable) the likelihood ratio test also favored the simpler models.

We calculated a standardized mean difference (SMD) by dividing the difference between the placebo and drug change scores by the pooled standard deviation of the change score at endpoint (imputed where necessary).

For PD, we applied a generalized linear (negative binomial) mixed model, using the glmer.nb command from the lme4 package (version 1.1-12). The effect measure of interest was the number of panic attacks per two weeks. Because this measure was highly skewed, we replaced values higher than 100 (45 (0.4%) of 11,785 observations) by a new value between 70 and 100 (randomly drawn from a uniform distribution) to improve the distribution and model convergence.

The initial model included the same fixed effects as for the other disorders. However, due to convergence problems, time was centered at the mean rather than at endpoint. For the covariance structure, we considered only a random trial-level intercept, and a random intercept and random effect for linear time at the participant level, as models including a random effect for quadratic time failed to converge. Since the lme4 package does not easily allow for either autocorrelation terms or various covariance structures, we

only modeled an unstructured covariance matrix.

We subsequently selected the best-fitting model using backward selection of the fixed effects, as done for the other disorders. Because of the non-normal distribution of panic attacks, we only calculated the endpoint scores and did not include a standardized difference for PD.

For all disorders, we also analyzed models that included age and gender as covariates. These yielded similar results as models without age and gender and are not described further.

Results

Trials and participants

We identified 34 trials, but we excluded one trial of paroxetine for PD [140] a priori, as it did not distinguish between full and limited-symptom panic attacks. We were denied access to 4 other trials: electronic data was not available for a trial of fluoxetine for OCD [166] (completed in 1991), while the translation costs for three Japanese trials of paroxetine for SAD [444, 445] and GAD [446] were considered prohibitive. One of these trials was positive, i.e. had statistically significant results for the primary outcome, while the other three were negative.

We received access to 29 trials with 3,656 placebo-treated and 5,323 antidepressant-treated participants. For GAD, we had access to 8 trials (6 positive) with 1,342 placebo-treated and 2,088 antidepressant-treated participants; for SAD, 4 trials (all positive) with 514 placebo-treated and 681 antidepressant-treated participants; for OCD, 4 trials (3 positive) with 350 placebo-treated and 782 antidepressant-treated participants; for PTSD, 3 trials (2 positive) with 459 placebo-treated and 612 antidepressant-treated participants; and for PD, 10 trials (5 positive) with 991 placebo-treated and 1,160 antidepressant-treated participants (see Table 10.1 for baseline characteristics and Table 10.4 in the Appendix for individual trial information).

GAD, SAD, OCD, and PTSD

For all four disorders, a model with an unstructured covariance matrix (including all random effects) and autocorrelated errors fit best. For GAD, the best-fitting model included the two-way interaction between baseline score and treatment group, but not the three-way interactions between baseline score, treatment group, and time (Figure 10.1a).

Table 10.1: *Baseline characteristics for each disorder*

	Female	Mean age	Baseline score		
Disorder	(%)	(SD)	Mean (SD)	Median (IQR)	Range
GAD	62.3	42.0 (13.4)	25.1 (5.9)		2 – 50
SAD	46.8	37.3 (11.0)	80.2 (24.0)		7 – 139
OCD	44.4	38.7 (12.4)	25.0 (5.3)		10 – 40
PTSD	62.6	41.1 (11.7)	75.2 (16.6)		30 – 132
PD	61.4	37.3 (10.5)		5 (3 – 11)	0 – 99

GAD: generalized anxiety disorder; OCD: obsessive-compulsive disorder; SAD: social anxiety disorder; PD: panic disorder; PTSD: post-traumatic stress disorder. The baseline score is based on the Hamilton Anxiety Rating Scale (HAM-A) for GAD, the Liebowitz Social Anxiety Scale (LSAS) for SAD, the Yale-Brown Obsessive-Compulsive Scale (Y-BOCS) for OCD, the Clinician-Administered PTSD Scale (CAPS) for PTSD, and the number of panic attacks/2 weeks for panic disorder.

For SAD, OCD and PTSD, the best-fitting model did not include any of the interactions between baseline score and treatment group (Figure 10.1b-d). Model specifications are available in Tables 10.5 and 10.6 in the Appendix.

For GAD, the estimated benefit of antidepressants (compared to placebo) at trial endpoint (8 weeks) was 1.4 (95% CI: 0.4-2.5, SMD: 0.15) points on the HAM-A for participants with a baseline score of 10, increasing to 4.0 (95% CI: 3.4-4.6, SMD: 0.43) for participants with a baseline score of 30 (see Table 10.2).

For SAD, OCD, and PTSD, the estimated benefit of antidepressants was the same across the severity range. For SAD, it was 16.1 (95% CI: 12.9-19.3, SMD: 0.59) points on the LSAS at week 12; for OCD, it was 3.4 (95% CI: 2.5-4.4, SMD: 0.39) points on the Y-BOCS at week 12; and for PTSD it was 10.3 (95% CI: 6.9-13.6, SMD: 0.41) points on the CAPS at week 12.

Panic disorder

For PD, the model with the lowest AIC (38267.2) contained the two-way interaction between baseline severity and group, but because this term was not significant ($p = 0.11$), we preferred a more parsimonious model without the interaction and with an only marginally larger AIC (38268.6 after removing two non-significant terms). Full specifications are provided in Tables 10.5 and 10.6 in the Appendix.

This parsimonious model indicated that the drug-placebo difference was constant on the log scale of the negative binomial model and hence that the *ratio* of the endpoint number of panic attacks/2 weeks in the placebo group compared to the drug group was constant (2.46) on the original scale. Consequently, the absolute difference between the drug and

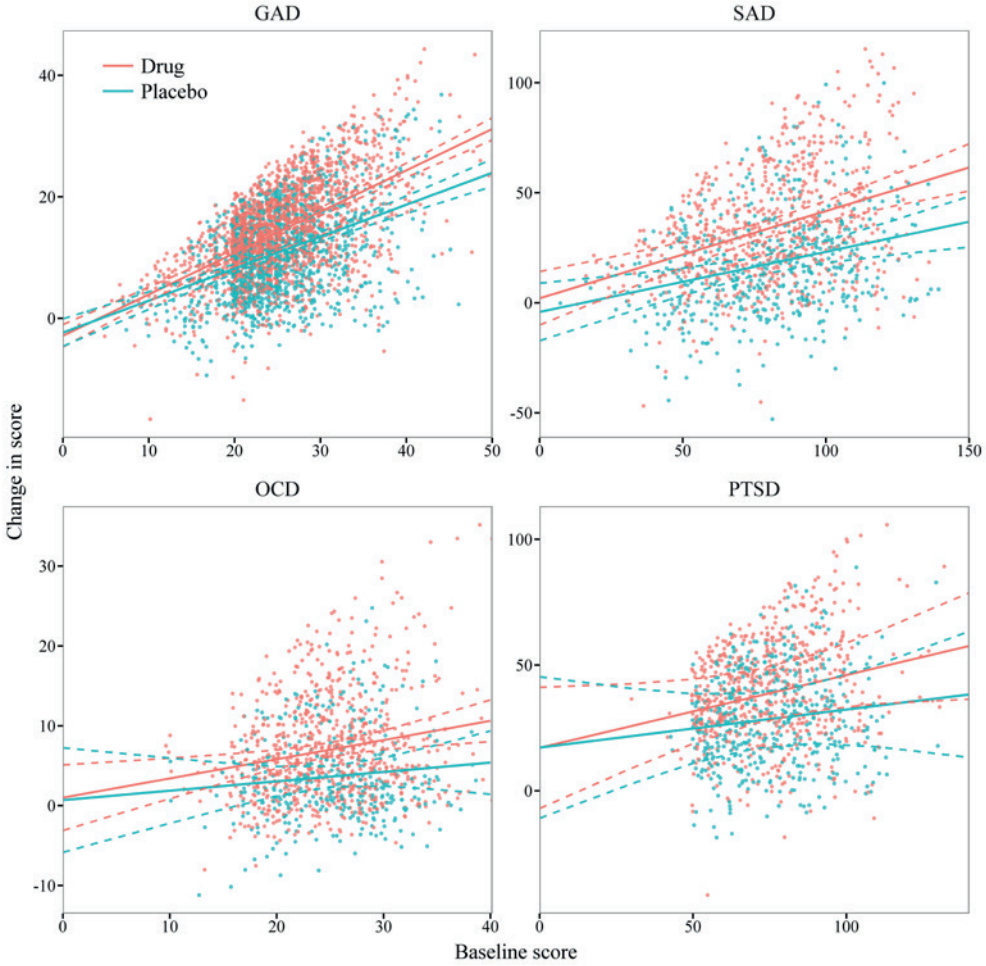

Figure 10.1: *Predicted change from baseline for antidepressant- and placebo-treated participants with generalized anxiety disorder (GAD), social anxiety disorder (SAD), obsessive-compulsive disorder (OCD), or post-traumatic stress disorder (PTSD). Predictions are derived from the full model, including non-significant interaction terms. Data points were jittered to reduce over-plotting.*

placebo groups actually increased with increasing severity (Figure 10.2).

For participants experiencing two panic attacks/2 weeks at baseline, the estimated drug-placebo difference was 0.2 (95% CI: 0.2-0.3) (in favor of antidepressants) at week 10. This increased to 0.4 (95% CI: 0.3-0.6) for participants experiencing 10 panic attacks/2 weeks at baseline, 0.9 (95% CI: 0.7-1.2) for participants experiencing 20 panic attacks/2 weeks at baseline, and 4.7 (95% CI: 3.0-6.4) for participants experiencing 40 panic attacks/2 weeks at baseline (Table 10.3).

Table 10.2: *Predicted change on the Hamilton Anxiety Rating Scale (HAM-A) and antidepressant-placebo difference after 8 weeks of treatment for GAD.*

		Predicted change (95% CI)			
Baseline	N	Placebo	Antidepressant	Diff. (95% CI)	SMD
10	79	2.7 (1.7 – 3.7)	4.1 (3.2 – 5.0)	1.4 (0.4 – 2.5)	0.15
15	259	5.4 (4.6 – 6.2)	7.4 (6.8 – 8.1)	2.1 (1.3 – 2.9)	0.22
20	1388	8.1 (7.5 – 8.7)	10.8 (10.3 – 11.3)	2.7 (2.1 – 3.3)	0.29
25	998	10.8 (10.3 – 11.3)	14.2 (13.7 – 14.6)	3.3 (2.8 – 3.9)	0.36
30	442	13.5 (12.9 – 14.1)	17.5 (17.0 – 18.0)	4.0 (3.4 – 4.6)	0.43
35	187	16.2 (15.5 – 17.0)	20.9 (20.2 – 21.5)	4.6 (3.8 – 5.4)	0.50
40	48	18.9 (17.9 – 20.0)	24.2 (23.3 – 25.1)	5.3 (4.2 – 6.3)	0.56
45	10	21.7 (20.4 – 22.9)	27.6 (26.5 – 28.7)	5.9 (4.6 – 7.2)	0.63

Baseline indicates the baseline score on the Hamilton Anxiety Rating Scale (HAM-A). N indicates the number of participants with a baseline score in between the indicated score and the subsequent score (e.g. 10 includes participants with baseline scores between 10 and 14) or the maximum score. Diff.: drug-placebo difference; GAD: generalized anxiety disorder; SMD: standardized mean difference.

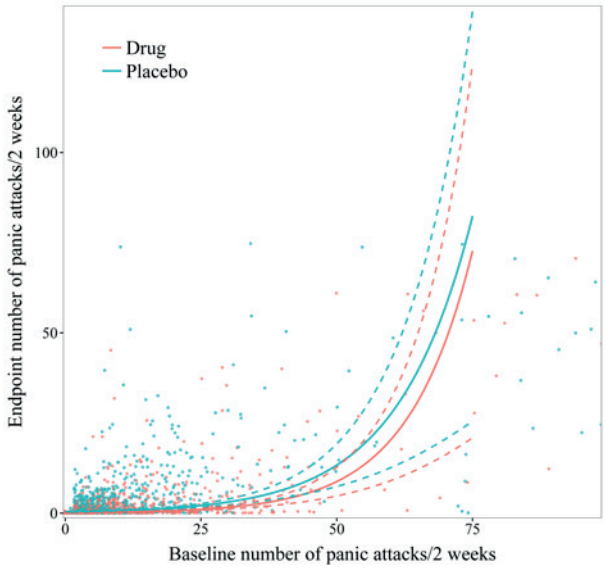

Figure 10.2: *Predicted number of panic attacks/2 weeks at endpoint for antidepressant- and placebo-treated participants with panic disorder. Predictions are derived from the full model, including non-significant interaction terms. Data points were jittered to reduce over-plotting.*

Table 10.3: *Predicted endpoint number of panic attacks/2 weeks and antidepressant-placebo difference after 10 weeks of treatment for panic disorder.*

		Predicted no. PAs/2 weeks (95% CI)		
Baseline	N	Placebo	Antidepressant	Diff. (95% CI)
2	636	0.38 (0.31 – 0.46)	0.15 (0.13 – 0.19)	0.23 (0.16 – 0.29)
4	396	0.44 (0.37 – 0.54)	0.18 (0.15 – 0.22)	0.26 (0.19 – 0.34)
6	258	0.52 (0.43 – 0.63)	0.21 (0.17 – 0.26)	0.31 (0.22 – 0.40)
8	175	0.61 (0.51 – 0.74)	0.25 (0.21 – 0.30)	0.36 (0.26 – 0.47)
10	240	0.72 (0.60 – 0.86)	0.29 (0.24 – 0.35)	0.43 (0.30 – 0.55)
15	123	1.07 (0.89 – 1.29)	0.43 (0.36 – 0.52)	0.64 (0.45 – 0.82)
20	133	1.59 (1.31 – 1.93)	0.65 (0.53 – 0.78)	0.95 (0.67 – 1.23)
30	58	3.54 (2.82 – 4.45)	1.44 (1.14 – 1.80)	2.10 (1.43 – 2.78)
40	44	7.86 (5.93 – 10.41)	3.19 (2.41 – 4.21)	4.67 (2.98 – 6.37)
60	37	38.76 (25.78 – 58.28)	15.72 (10.50 – 23.55)	23.04 (12.23 – 33.83)

Baseline indicates the baseline number of panic attacks (PAs)/2 weeks. N indicates the number of participants with a baseline score in between the indicated score and the subsequent score (e.g. 2 includes participants with baseline scores of 2 or 3) or the maximum score (99). Diff.: drug-placebo difference.

Discussion

Principal findings

This is the first individual patient data meta-analysis examining the relationship between baseline severity and antidepressant efficacy for anxiety disorders. We showed that initial severity moderates antidepressant efficacy for GAD, but not for SAD, OCD, and PTSD. For PD, the ratio between the number of panic attacks in the placebo group compared to the drug group was constant, but the absolute difference between antidepressants and placebo was small for patients experiencing few panic attacks at baseline and increased with increasing severity. For all disorders, a regression to the mean effect occurred, but this cannot explain the interaction between baseline severity and treatment in GAD.

Our findings are in agreement with our earlier trial-level meta-analysis for SAD, OCD, and PTSD, but not for GAD and PD [433]. These differences are likely because of the much larger sample size and the use of IPD in this study. The SMDs for SAD and PTSD were also larger than those found earlier [20], which is probably due in part to trial selection, as the paroxetine trials had higher effect sizes than trials of other drugs for these disorders, and in part to different analytical techniques.

Because the non-normal primary outcome necessitated an alternative analytical approach, our findings for PD are difficult to compare to the other disorders. However, it is interesting that we found a relationship between initial severity and antidepressant efficacy for GAD, but not for SAD, OCD, and PTSD.

GAD is often considered to be more closely related to MDD than the other anxiety disorders. In factor analyses, GAD often clusters with depression in a 'distress' dimension while other anxiety disorders cluster in a 'fear' dimension [5], although PTSD may also load primarily on the 'distress' dimension [447, 448]. Additionally, the HAM-A overlaps with the Hamilton Depression Rating Scale (HAM-D) commonly used in MDD trials.

On the other hand, since the association between initial severity and antidepressant efficacy in MDD has been called into question [75, 76], a greater similarity between MDD and GAD might not explain our findings. HAM-A items also tend to be relatively non-specific, covering such common symptoms as insomnia, tension, worries, and pains, while the LSAS and Y-BOCS questionnaires specifically examine the distress associated with respectively feared social situations and obsessions or compulsions, and the CAPS questionnaire examines both general distress symptoms and specific trauma-related distress. Such general distress symptoms, particularly when mild, may be more responsive to placebo or more likely to improve spontaneously, which could explain why antidepressants provide little benefit over placebo in mild GAD.

However, it is also important to note that we had the largest sample size for GAD. Although we had more than 1000 participants for each disorder and hence should have been able to detect a substantial interaction effect if it existed, it is possible that smaller interaction effects for the other disorders were missed. However, these are less likely to be of clinical significance.

Strengths and limitations

The main strength of this study is that we used IPD and had a large sample size for each disorder. Furthermore, we used disorder-specific primary outcomes and made full use of the longitudinal data by employing mixed models.

Our study is limited by the limitations of the included trials. In particular, minimum severity criteria restricted the number of participants at the low end of the severity range for some disorders. Half of the GAD trials specified a minimum HAM-A score of 20, for instance, even though most primary care patients diagnosed with GAD have scores lower than 20 [449]. Hence, our findings are most clearly applicable to patients who are moderately or severely ill and less so to patients with subthreshold or very mild symptoms. The included trials also frequently excluded patients with comorbid disorders such as MDD, even though these disorders commonly occur together [4].

Furthermore, our findings for PD are difficult to compare to the other disorders. The best-fitting model showed that the ratio of the number of panic attacks at endpoint in the placebo group compared to the drug group remained constant, but this means that the drug-placebo difference increased with increasing severity. We have emphasized the latter, because this measure is most comparable to the other disorders, but other choices could

be made. Additionally, while the number of panic attacks is a clinically relevant outcome, other important facets of PD, such as agoraphobia, were not examined. Our results may have been different if we had been able to use a more comprehensive questionnaire, such as the Panic Disorder Severity Scale [450].

We also did not receive data for four trials. Since three of these trials were negative, we may have overestimated the antidepressant effect for GAD, SAD, and OCD, but it seems unlikely that this would have affected our findings regarding initial severity. Negative trials will probably show little evidence for differential efficacy, so it is not likely that we would have found a significant interaction effect for SAD and OCD if we had been able to include participants from these trials. For GAD, the evidence in favor of an interaction effect was sufficiently strong that it probably would have remained even if we had added an additional trial in which differential efficacy was not apparent.

Finally, we included only trials of duloxetine, paroxetine, and fluoxetine. We used Clinical Study Data Request because it allowed us to obtain a nearly complete set of trials for these drugs. Other approaches (e.g., a comprehensive literature search followed by requesting IPD from authors) would almost certainly have introduced much more significant biases into our trial selection, because of reporting bias [20] and refusal or inability to share data. For example, a study that took this approach for MDD trials only received data for 6 of 23 eligible trials [72]. However, future research should examine other antidepressants.

Clinical implications

To understand the implications of these findings, it should be noted that the clinical relevance of a treatment effect is context-specific, depending on such factors as the expected sequelae of the disease, the costs and drawbacks of the treatment, and the efficacy of alternative treatments [451]. Without an agreed-upon cut-off point for a clinically relevant effect, it is difficult to establish a threshold below which the effects of antidepressants for GAD and PD are not clinically meaningful. For GAD, the SMDs do suggest that the antidepressant-placebo difference is small for patients with a baseline severity score of 15 or less.

Even without a definite cut-off point, though, it is clear that the risk-benefit ratio for GAD and PD becomes less favorable as initial severity decreases. It is therefore imperative that clinicians transparently discuss the expected benefits of antidepressants with patients with mild to moderate symptoms, who constitute the majority of treatment-seeking patients in primary care [449]. To our knowledge, there is no evidence that alternative treatments, such as psychotherapy, would be more effective than antidepressants for patients with mild GAD or PD. These modalities may nevertheless be preferable as they are often thought to have fewer adverse effects (although poor monitoring limits our understanding of the negative effects of psychotherapy [47, 49]).

There was no evidence for a relationship between initial severity and antidepressant efficacy for SAD, OCD, and PTSD. Nevertheless, other factors, such as anticipated course, patient preferences, and the availability, acceptability, and efficacy of alternative treatments, could still lead to different prescribing decisions for mild versus severe disorders, even in the absence of differential efficacy.

Conclusions

We found that antidepressants are equally effective across the severity range for SAD, OCD, and PTSD. For GAD and PD, however, the benefits of antidepressants over and above placebo are small to negligible at low severity. The trade-off between benefits and risks may therefore be unfavorable for these patients, and alternative approaches that might have fewer risks, such as guided self-help or cognitive-behavioral therapy, may be preferred as first-line treatment.

Appendix

Table 10.4: *Supplemental table of studies*

Disorder	Drug	Trial	Dose (mg/day)	Duration (weeks)	Sample size		Baseline
					Drug	Placebo	Mean (SD)
GAD	Duloxetine	HMBR [134]	60, 120	9	334	173	25.3 (7.4)
		HMDT [135]	60 – 120	10	161	158	23.0 (7.7)
		HMDW [452]	60 – 120 V 75 – 225	10	392	163	27.5 (7.5)
		HMDU [136]	60 – 120 V 75 – 225	10	308	158	25.2 (5.7)
	Paroxetine	637 [126]	20 – 50	8	184	184	25.6 (4.5)
		641 [132]	20, 40	8	385	180	24.1 (3.6)
		642 [133]	20 – 50	8	161	163	24.1 (3.6)
	Paroxetine CR	791 [453]	12.5 – 37.5	8	163	163	24.6 (3.7)
SAD	Paroxetine	382 [152]	20 – 50	12	90	92	80.3 (23.5)
		454 [153]	20, 40, 60	12	275	93	77.3 (23.0)
		502 [151]	20 – 50	12	131	145	86.4 (24.5)
	Paroxetine CR	790 [154]	12.5 – 37.5	12	185	184	78.5 (24.0)
OCD	Fluoxetine	HCEP [165]	20, 40, 60	13	259	88	24.1 (5.4)
	Paroxetine	116 [169]	20, 40, 60	12	250	88	25.4 (5.2)
		118 [454]	20 – 60	12	79	75	24.2 (4.8)
		136 [170]	20 – 60	12	194	99	25.9 (5.2)
PTSD	Paroxetine	627 [127]	20 – 50	12	154	159	77.9 (17.9)
		648 [164]	20 – 50	12	136	133	73.4 (16.1)
		651 [163]	20, 40	12	322	167	74.5 (15.8)
Panic disorder	Fluoxetine	HCJC [429]	20 – 60	12	90	90	7.8 (6.6)
		HCJB (N/A)	20 – 60	12	106	104	6.8 (7.1)
		HCHG (N/A)	10, 20	10	145	63	10.7 (16.4)
		HCHQ (N/A)	20	8	56	65	22.4 (41.9)
	Paroxetine	120 [139]	10, 20, 40	10	182	66	10.2 (15.9)
		187 [141]	20 – 60	12	109	113	12.5 (13.8)
		223 [455]	10 – 60	10	68	67	8.4 (9.2)
	Paroxetine CR	494 [142]	25 – 75	10	127	137	10.4 (20.9)
		495 [142]	25 – 75	10	143	155	10.0 (12.7)
		497 [142]	25 – 75	10	134	131	8.8 (11.8)

N/A indicates unpublished trials for which a registry summary could not be located. These trials were identified in the Food and Drug Administration drug application package for fluoxetine for panic disorder. CR: controlled release; GAD: generalized anxiety disorder; OCD: obsessive-compulsive disorder; PTSD: post-traumatic stress disorder; SAD: social anxiety disorder; V: venlafaxine (used as an active comparator).

Table 10.5: *Model comparisons*

Disorder	Model	Specifications	DF	AIC	BIC	Log-likelihood or deviance (PD only)
GAD	1	Full model	21	32046.47	32209.12	-16002.23
	2	Model 1 - $B \times G \times T^2$	20	32044.61	32199.52	-16002.31
	3	Model 2 - $B \times G \times T$	19	32043.03	32190.19	-16002.51
	4	Model 3 - $B \times G$	18	32058.99	32198.41	-16011.49
SAD	1	Full model	21	10633.50	10778.63	-5295.749
	2	Model 1 - $B \times G \times T^2$	20	10632.75	10770.96	-5296.372
	3	Model 2 - $B \times G \times T$	19	10631.48	10762.79	-5296.741
	4	Model 3 - $B \times G$	18	10631.06	10755.45	-5297.528
	5	Model 4 - $B \times T^2$	17	10637.17	10754.65	-5301.583
OCD	1	Full model	21	13500.48	13646.47	-6729.238
	2	Model 1 - $B \times G \times T^2$	20	13498.79	13637.84	-6729.396
	3	Model 2 - $B \times G \times T$	19	13498.10	13630.19	-6730.050
	4	Model 3 - $B \times G$	18	13496.15	13621.29	-6730.076
	5	Model 4 - $B \times T^2$	17	13498.81	13617.00	-6732.405
PTSD	1	Full model	21	7152.29	7281.26	-3555.145
	2	Model 1 - $B \times G \times T^2$	20	7150.42	7273.24	-3555.210
	3	Model 2 - $G \times T^2$	19	7149.29	7265.97	-3555.643
	4	Model 3 - $B \times G \times T$	18	7148.57	7259.12	-3556.285
	5	Model 4 - $B \times G$	17	7146.98	7251.38	-3556.489
	6	Model 5 - $B \times T^2$	16	7146.47	7244.74	-3557.235
	7	Model 6 – $B \times T$	15	7149.19	7241.32	-3559.597
PD	1	Full model	17	38269.0	38390.9	38235.0
	2	Model 1 - $B \times G \times T^2$	16	38269.1	38383.9	38237.1
	3	Model 2 - $G \times T^2$	15	38267.8	38375.4	38237.8
	4	Model 3 - $B \times G \times T$	14	38267.2	38367.6	38239.2
	5	Model 4 - $B \times G$	13	38267.8	38361.0	38241.8
	6	Model 5 - $B \times T^2$	12	38268.6	38354.6	38244.6
	7	Model 6 - T^2	11	38274.7	38353.6	38252.7

B: baseline score; G: group (drug vs. placebo); T: time (in days since baseline); T^2: quadratic time. AIC: Akaike information criterion; BIC: Bayes information criterion; DF: degrees of freedom; GAD: generalized anxiety disorder; OCD: obsessive-compulsive disorder; PD: panic disorder; PTSD: post-traumatic stress disorder; SAD: social anxiety disorder.

Table 10.6: *Model specifications*

		Full model			Best-fitting model		
Disorder	Parameter	β	SE	p-value	β	SE	p-value
GAD	Intercept	0.189	0.033	<0.001	0.189	0.033	<0.001
	T	0.015	0.018	0.390	0.015	0.018	0.390
	T^2	-0.127	0.008	<0.001	-0.127	0.008	<0.001
	G	0.420	0.035	<0.001	0.420	0.035	<0.001
	B	0.391	0.027	<0.001	0.403	0.022	<0.001
	$T \times G$	0.019	0.023	<0.001	0.019	0.023	0.405
	$T^2 \times G$	-0.049	0.010	<0.001	-0.049	0.010	<0.001
	$T \times B$	0.0383	0.016	0.019	0.039	0.010	<0.001
	$T^2 \times B$	-0.023	0.007	0.002	-0.025	0.004	<0.001
	$B \times G$	0.115	0.035	<0.001	0.095	0.022	<0.001
	$B \times G \times T$	0.002	0.021	0.941	-	-	-
	$B \times G \times T^2$	-0.004	0.009	0.706	-	-	-
SAD	Intercept	0.164	0.055	0.003	0.163	0.055	0.003
	T	-0.012	0.028	0.665	-0.012	0.028	0.672
	T^2	-0.074	0.010	<0.001	-0.074	0.010	<0.001
	G	0.710	0.071	<0.001	0.711	0.071	<0.001
	B	0.289	0.052	<0.001	0.359	0.035	<0.001
	$T \times G$	0.134	0.037	<0.001	0.134	0.037	<0.001
	$T^2 \times G$	-0.033	0.013	0.008	-0.034	0.013	0.008
	$T \times B$	0.049	0.028	0.076	0.059	0.019	0.001
	$T^2 \times B$	-0.014	0.009	0.120	-0.016	0.006	0.010
	$B \times G$	0.130	0.070	0.065	-	-	-
	$B \times G \times T$	0.018	0.037	0.624	-	-	-
	$B \times G \times T^2$	-0.004	0.013	0.770	-	-	-
OCD	Intercept	-0.058	0.067	0.380	-0.063	0.067	0.346
	T	-0.045	0.044	0.304	-0.045	0.044	0.305
	T^2	-0.066	0.015	<0.001	-0.066	0.015	<0.001
	G	0.559	0.080	<0.001	0.562	0.080	<0.001
	B	0.102	0.069	0.142	0.178	0.037	<0.001
	$T \times G$	0.072	0.052	0.163	0.072	0.052	0.166
	$T^2 \times G$	-0.046	0.018	0.011	-0.047	0.018	0.010
	$T \times B$	-0.010	0.046	0.833	-0.006	0.024	0.813
	$T^2 \times B$	-0.010	0.016	0.520	-0.018	0.008	0.034
	$B \times G$	0.107	0.082	0.195	-	-	-
	$B \times G \times T$	0.005	0.054	0.928	-	-	-
	$B \times G \times T^2$	-0.011	0.019	0.552	-	-	-
PTSD	Intercept	0.127	0.104	0.220	0.129	0.104	0.216
	T	0.055	0.047	0.241	0.025	0.035	0.481
	T^2	-0.072	0.017	<0.001	-0.084	0.012	<0.001
	G	0.423	0.071	<0.001	0.425	0.071	<0.001
	B	0.103	0.056	0.062	0.159	0.035	<0.001
	$T \times G$	0.016	0.062	0.796	0.068	0.025	0.006

continued

Table 10.6: *Model specifications*

		Full model			Best-fitting model		
Disorder	Parameter	β	SE	p-value	β	SE	p-value
	$T^2 \times G$	-0.020	0.022	0.355	-	-	-
	$T \times B$	-0.011	0.047	0.816	-0.026	0.012	0.034
	$T^2 \times B$	-0.008	0.016	0.610	-	-	-
	$B \times G$	0.095	0.072	0.185	-	-	-
	$B \times G \times T$	0.006	0.061	0.923	-	-	-
	$B \times G \times T^2$	-0.008	0.021	0.721	-	-	-
Panic disorder	Intercept	0.446	0.070	<0.001	0.435	0.069	<0.001
	T	-0.671	0.030	<0.001	-0.673	0.068	<0.001
	T^2	0.027	0.030	0.084	0.033	0.011	0.004
	G	-0.493	0.071	<0.001	-0.464	0.068	<0.001
	B	0.846	0.045	<0.001	0.903	0.032	<0.001
	$T \times G$	-0.357	0.040	<0.001	-0.362	0.039	<0.001
	$T^2 \times G$	0.027	0.023	0.241	-	-	-
	$T \times B$	0.092	0.024	<0.001	0.114	0.018	<0.001
	$T^2 \times B$	-0.004	0.011	0.722	-	-	-
	$B \times G$	0.144	0.065	0.026	-	-	-
	$B \times G \times T$	0.037	0.035	0.295	-	-	-
	$B \times G \times T^2$	-0.024	0.017	0.148	-	-	-

B: baseline score; G: group (drug vs. placebo); T: time (in days since baseline); T^2: quadratic time. GAD: generalized anxiety disorder; OCD: obsessive-compulsive disorder; PD: panic disorder; PTSD: post-traumatic stress disorder; SAD: social anxiety disorder.

Chapter 11

Early improvement in depressive symptoms and response to antidepressants: an individual patient data meta-analysis

Ymkje Anna de Vries, Annelieke M. Roest, Elske H. Bos,
J. G. M. (Hans) Burgerhof, Hanna M. van Loo, Peter de Jonge

Submitted

Abstract

Objective: To investigate whether early improvement of individual depressive symptoms during antidepressant treatment predicts response or remission.

Method: We obtained individual patient data of 2,184 placebo-treated and 6,058 antidepressant-treated participants from 30 trials for major depressive disorder (MDD). The primary outcome was response on the Hamilton Depression Rating Scale (HAM-D) at week 6; secondary outcomes were remission at week 6 and response and remission at week 12. Least absolute shrinkage and selection operator (lasso) logistic regression was used for variable selection. We compared models that only included early improvement in the total HAM-D score by week 2 ($\geq$ 20% improvement vs. <20% improvement) (total improvement model) with models that included early improvement in individual HAM-D items (item improvement model) and models with interactions among early-improving HAM-D items, age, and gender (item interactions model).

Results: The models that included individual items performed slightly, but significantly, better than the total improvement model. By week 6, 51% of all antidepressant-treated participants responded, but participants who were predicted not to respond had a 29% chance of response. By week 12, the overall probability of response was 69%; of participants who were predicted not to respond, 43% responded. In post-hoc analyses with early improvement as a continuous variable, including individual items did not enhance model performance.

Conclusions: Examining individual symptoms does not add meaningfully to the predictive ability of early improvement in the total score, particularly if improvement is seen as continuous. In addition, absence of early improvement does not rule out good outcomes. Therefore, adapting treatment because of limited improvement by week two is not supported by our findings.

Introduction

Antidepressants are first-line treatments for major depressive disorder (MDD) [73, 74, 86]. However, many patients fail to respond, with response rates averaging around 50% in clinical trials [59], and it is important to identify these patients as soon as possible to minimize the duration of ineffective treatment and the time until response.

Clinical guidelines currently recommend 4 - 8 weeks of treatment before evaluating the effects of treatment and considering a change in management for patients who show no improvement [73, 74, 86]. The evidence base for this recommendation, however, is limited. At a group level, antidepressant effects can be detected within the first week of treatment [456], and at the level of an individual patient numerous studies have found that improvement within the first one or two weeks of treatment is associated with later response or remission [87, 88, 94, 95, 457, 458, 459, 460, 461].

Despite this general consensus, these studies disagree on whether lack of early improvement is a sufficiently good predictor to justify a change in management. For instance, one study found that only 4% of participants with no improvement after two weeks reached remission by week four [87], suggesting that non-improvers have virtually no chance of good outcomes. In a different study that followed participants over a longer period of time, on the other hand, 44% of participants without early improvement still responded after twelve weeks of treatment [95].

On average, most studies indicate that at least 20% to 30% of participants without early improvement attain a good outcome after four to twelve weeks of treatment, which is reduced from the overall probability of around 50% but not negligible [461]. Conversely, many early-improvers do not achieve good outcomes. Hence, better predictive models are desirable.

One possibility to extend these models is to examine individual symptoms, rather than only the total depression score. There are meaningful differences between symptoms (e.g. regarding risk factors and disability) [96], and severity of specific symptoms has been found to be associated with prognosis [83, 462]. Previous studies have also found that response or remission can be predicted by early improvement in several specific symptoms, including depressed mood, somatic symptoms, loss of insight, and others [98, 99, 100, 463, 464, 465].

However, these studies did not investigate whether improvement in individual symptoms is more informative than improvement in the total score alone. In the current study, we therefore investigated this question. We also examined whether there are interactions between early-improving symptoms, gender, and age. Finally, we examined whether individual symptoms are differentially predictive for response to different antidepressant classes.

Methods

Data source and trial selection

We requested individual patient data (IPD) from Clinical Study Data Request [441], a data-sharing platform providing data from (among others) trials of antidepressants developed by sponsors using this platform (GlaxoSmithKline and Lilly).

We examined second-generation antidepressants (SGAs, defined as selective serotonin reuptake inhibitors (SSRIs), serotonin-norepinephrine reuptake inhibitors (SNRIs), or other antidepressants approved after 1987), since older antidepressants are considered second-line options. However, we also included trials of new chemical entities, which were never approved for MDD, if an approved SGA was used as an active comparator.

Trials were required to be randomized, placebo- or active-comparator-controlled, and double-blind, and to have a minimum duration of 6 weeks, with trial visits in which the Hamilton Depression Rating Scale (HAM-D) was administered at baseline, week 2, and either week 6 (± 1) or week 12 (± 1) (or both). We excluded trials in children (<18 years old), trials for non-MDD indications, and trials that specifically included only participants with additional symptoms (e.g. MDD with pain).

Patient population

We only included participants assigned to placebo or SGAs. No eligible trials included participants assigned to non-SSRI/SNRI SGAs (e.g. mirtazapine), so our final sample consisted of participants assigned to placebo, SSRIs, or SNRIs.

We took a complete-case approach, only including participants who had valid HAM-D scores at baseline, week 2, and week 6 or 12. Week 2 visits took place on day 14 (± 7 days), week 6 visits on day 42 (± 14 days), and week 12 visits on day 84 (± 14) days. If a participant had multiple visits within the eligible time frame, we selected the visit closest to the intended visit day or, if eligible visits were equally close to the intended visit day (e.g. day 35 and day 49), we randomly selected one of the visits.

Training and test data

We randomly split the data into an 80% training set and 20% test set, stratified by treatment group (placebo, SSRI, or SNRI). The training set was used for model discovery and cross-validation, while prediction accuracy was assessed in the test set.

Outcomes and predictors

Our primary outcome was response ($\geq$ 50% reduction in HAM-D (17-item version) score) at week 6 [466]. Secondary outcomes were remission (score of $\leq$ 7 on the HAM-D-17) at week 6, and response and remission at week 12.

Improvement in symptoms was calculated from the baseline and week 2 HAM-D items and dichotomized into "no improvement" (no change or worsening) and "improvement" (improvement in the item score of $\geq$ 1). Baseline HAM-D items were dichotomized into absent (score of 0) or present (score of 1). Early improvement on the total HAM-D score was dichotomized into no improvement (<20% improvement) or improvement ($\geq$ 20% improvement), consistent with other studies [461], while the baseline HAM-D score was standardized. As demographic variables, we included age (standardized) and gender.

Statistical analysis

Our primary analyses only included antidepressant-treated participants. For variable selection, we used least absolute shrinkage and selection operator (lasso) logistic regression [467], implemented in the glmnet package (version 2.0-5) for R (version 3.3.0).

For each outcome (response and remission at week 6 and 12), we built four models: (1) the baseline model, which included only baseline HAM-D score, HAM-D items, age, and gender; (2) a total improvement model, which included these baseline variables and early improvement in the total HAM-D score; (3) an item improvement model, which included all these variables and early improvement in the 17 HAM-D items; and (4) an item interactions model, which included all of the above and all two-way interactions among early-improving items, age, and gender.

The optimal regularization penalty (λ) was determined by ten-fold cross-validation. We favored sparser models by choosing the largest λ whose deviance was within one standard error of the minimal deviance [468]. From each lasso model, we selected all variables with non-zero coefficients to build a mixed-effects logistic regression model with a random intercept for trial, using the lme4 package (version 1.1-12). Hence, we built four separate mixed-effects models for each outcome.

Model performance

The prediction accuracy of each mixed-effects model was assessed in the test set by determining the area under the receiver operating characteristic (ROC) curve (AUC) (R package pROC, version 1.8). The model with the highest AUC was considered the best model. We also determined the accuracy, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) for each model by assigning participants with

a model-predicted probability of response/remission of $\geq 50\%$ to the response/remission group.

Secondary analyses and post-hoc analyses

We performed secondary analyses in the total group of both antidepressant- and placebo-treated participants. In these analyses, we included treatment group (placebo vs. SSRI vs. SNRI) as a predictor in the lasso regressions to examine whether associations between early-improving items and outcome were dependent on antidepressant class (suggesting a drug-specific mechanism).

In our main and secondary analyses, we dichotomized early improvement, for comparability with other studies. However, we conducted additional post-hoc analyses in which baseline item scores and early improvement (change from baseline in the total score and the individual items) were included as continuous variables.

Results

Trials and patients

We requested and received data for 32 trials. However, 2 trials proved to be ineligible (no week 6 or 12 visit). The remaining 30 trials investigated duloxetine (15 trials), paroxetine (13 trials), or new chemical entities (2 trials). Thirteen trials also included other SGAs (escitalopram, fluoxetine, paroxetine, or venlafaxine).

The total number of participants in these 30 trials was 10,365, of whom 8,242 participants had a week 6 visit. The ten trials with a duration of ≥ 12 weeks included 4,487 participants, of whom 3,103 had a week 12 visit. Sample characteristics are shown in Table 11.1. Table 11.3 in the Appendix provides further details about the individual trials.

Variable selection

Detailed information about the variables selected by the lasso regressions are provided in Tables 11.4 - 11.7 in the Appendix. In brief, all improvement models selected early improvement in the total score. The item improvement models generally selected most of the early-improving HAM-D items. However, items 3 (suicide) and 15 (hypochondria) were never selected, while items 1 (depressed mood), 2 (guilt), 4 (early insomnia), 7 (work and activities), 10 (psychological anxiety), and 13 (general somatic symptoms) were always selected. Baseline HAM-D items were selected infrequently. The item interaction models generally selected a number of interactions among early-improving symptoms;

Table 11.1: *Sample characteristics*

	Week 6 sample			Week 12 sample		
	Placebo	SSRI	SNRI	Placebo	SSRI	SNRI
Sample size	2,184	3,322	2,736	652	1,270	1,181
Mean baseline HAM-D (SD)	21.5 (5)	22.1 (4)	20.8 (5)	22.2 (5)	22.8 (4)	22.1 (5)
Mean age (SD)	44 (14)	43 (14)	45 (14)	50 (16)	45 (14)	48 (15)
Female (%)	63.5	61.8	65.7	64.7	62.7	65.6
Early improvement (%)	52.7	62.9	62.3	54.4	63.1	66.3
Response (%)	38.3	52.4	49.9	53.2	69.4	67.2
Remission (%)	22.6	32.1	32.7	34.5	49.4	48.9

HAM-D: Hamilton Depression Rating Scale; SNRI: serotonin-norepinephrine reuptake inhibitor; SSRI: selective serotonin reuptake inhibitor.

only for remission at week 12 were any interactions between symptoms and age or gender selected.

Model performance

ROC curves obtained from the mixed-effects logistic regression models for response at week 6 are shown in Figure 11.1. The baseline model performed quite poorly (AUC: 0.60). The total improvement model performed significantly better (AUC: 0.73), and the item improvement and item interactions model performed similarly (AUC: 0.77) and significantly better than the total improvement model. For remission at week 6 and response and remission at week 12, the patterns were similar, although model performance was worse for the week 12 outcomes (Table 11.2).

The accuracy, sensitivity, specificity, PPV, and NPV of each model are also given in Table 11.2. There were only minor differences between the three early improvement models. At week 6, 51% of antidepressant-treated participants in the test set responded and 33% remitted. The most parsimonious model with the highest AUC, the item improvement model, predicted non-response for 46% of participants; the associated NPV was 0.71, indicating that 29% of these participants were false negatives who did actually respond by week 6. Conversely, of participants who were predicted to respond, 70% actually responded (Figure 11.2). For remission at week 6, the model identified a large majority group (78% of participants) with a slightly reduced probability of remission (24%) and a minority group with an increased probability of remission (66%) (Figure 11.4 in the Appendix).

By week 12, 69% of participants in the test set responded and 51% remitted. The item improvement model predicted non-response for 16% of participants, but these participants

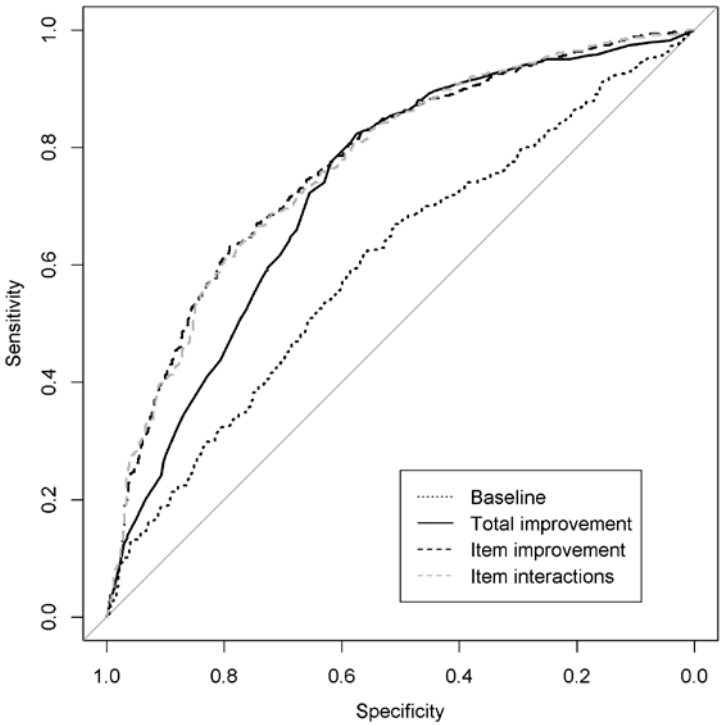

Figure 11.1: *Receiver-operating characteristic curve for the baseline, total improvement, item improvement, and item interactions model for response at week 6.*

still had a 43% probability of response. For remission, 47% of participants were predicted non-remitters, and these had a 34% probability of remission (see Figures 11.5 and 11.6 in the Appendix).

Post-hoc, we also used the predicted probability of responding or remitting to divide participants into quintiles and examined each quintile's actual probability of response or remission (Figures 11.7 - 11.10). This more fine-grained approach suggested that the improvement models could identify a risk group with poor outcomes at week 6, but prediction was less accurate and not much better than the baseline model by week 12.

Secondary analyses

Treatment group (placebo vs. SSRI vs. SNRI) was a significant predictor of response and remission at both week 6 and week 12. However, models that only included a main effect for treatment group performed as well as models with interactions between group and other variables (including gender, age, baseline score, baseline items, total improvement,

Table 11.2: *Model performance*

Week	Outcome	Model	AUC	Accuracy	Sensitivity	Specificity	PPV	NPV
6	Response	Baseline	0.60	0.59	0.67	0.51	0.59	0.59
		Total improvement	0.73	0.70	0.81	0.59	0.67	0.74
		Item improvement	0.77	0.70	0.74	0.66	0.70	0.71
		Item interactions	0.77	0.70	0.72	0.67	0.70	0.70
	Remission	Baseline	0.64	0.68	0.09	0.98	0.69	0.68
		Total improvement	0.74	0.71	0.30	0.92	0.64	0.72
		Item improvement	0.78	0.74	0.44	0.89	0.66	0.76
		Item interactions	0.78	0.74	0.43	0.89	0.66	0.76
12	Response	Baseline	0.62	0.69	1.00	0.00	0.69	N/A
		Total improvement	0.67	0.72	0.93	0.27	0.73	0.62
		Item improvement	0.71	0.71	0.90	0.29	0.73	0.57
		Item interactions	0.70	0.70	0.90	0.29	0.73	0.56
	Remission	Baseline	0.62	0.59	0.61	0.57	0.60	0.58
		Total improvement	0.68	0.64	0.69	0.59	0.64	0.64
		Item improvement	0.74	0.67	0.69	0.64	0.67	0.66
		Item interactions	0.73	0.66	0.65	0.68	0.68	0.65

AUC: Area under the (receiver operating characteristic) curve; NPV: negative predictive value; N/A: not applicable (undefined); PPV: positive predictive value.

and early-improving items), indicating no evidence for different associations between early-improving symptoms and response or remission depending on antidepressant class (Table 11.8 in the Appendix).

Post-hoc analyses

Because the total improvement model performed nearly as well as models that included individual items, we performed additional analyses using continuous change from baseline, since dichotomizing a continuous variable might affect model performance.

For response at week 6, the lasso regressions for the total improvement, item improvement, and item interaction models all selected the same variables (baseline score and change from baseline). The AUC of this model was 0.79. For remission at week 6 and response and remission at week 12, the lasso regressions did select individual items for the item improvement and/or the item interactions model. However, these models had identical or slightly (and non-significantly) worse AUCs than the (more parsimonious) total improvement model. The AUC for the total improvement model was 0.79 for remission at week 6; 0.71 for response at week 12; and 0.75 for remission at week 12.

Figure 3 depicts the probability of response or remission as a function of the percentage change from baseline in the total HAM-D score at week 2. The probability of response

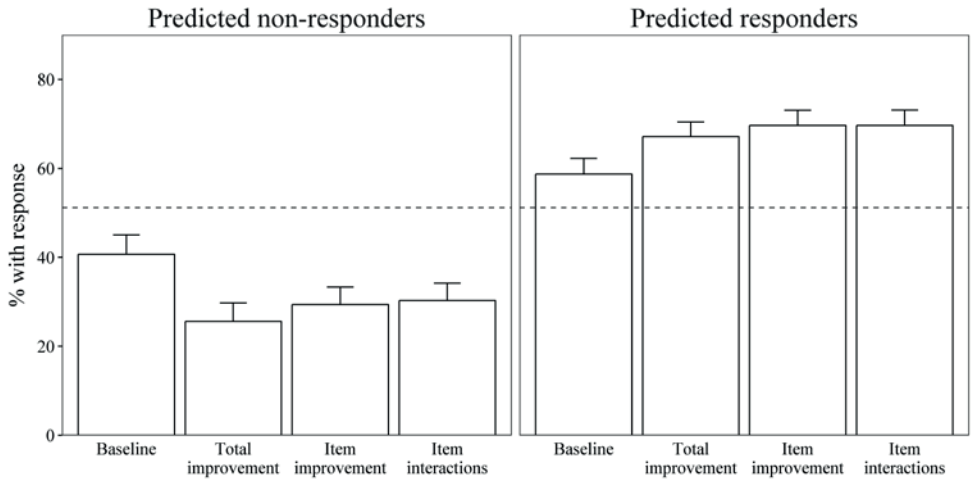

Figure 11.2: *Actual probability of response at week 6 according to participants' predicted outcome (non-response vs. response). The dashed line indicates the baseline probability of response. The models predicted non-response for 42% (baseline), 38% (total improvement), 46% (item improvement), and 47% (item interactions) of participants. Error bars indicate the 95% confidence interval.*

at week 6 was 91% for the few participants (163 (3%) of 6,058) who improved by 80% by week 2, decreasing to 17% for participants who showed any early worsening (573 (9%) participants). At week 12, however, even participants who showed early worsening still had a 39% probability of responding.

Discussion

In this individual patient data meta-analysis, we investigated whether early improvement in individual HAM-D symptoms could predict response and remission better than early improvement in the total score alone. Consistent with previous literature [461], we found that patients without early improvement were less likely to respond or remit. In our main analyses, a model with individual symptoms did perform better than a model that only included total improvement. However, the difference was relatively small, and secondary analyses examining continuous change from baseline did not confirm an added benefit from examining individual symptoms.

There was also no evidence that interactions between age, gender, and symptoms improved model performance. Our secondary analyses found no evidence that associations between early-improving symptoms and outcome differed between placebo, SSRIs, and SNRIs, since models that contained these interactions performed no better than models that did not. Taken together, our results show that early improvement is a non-specific

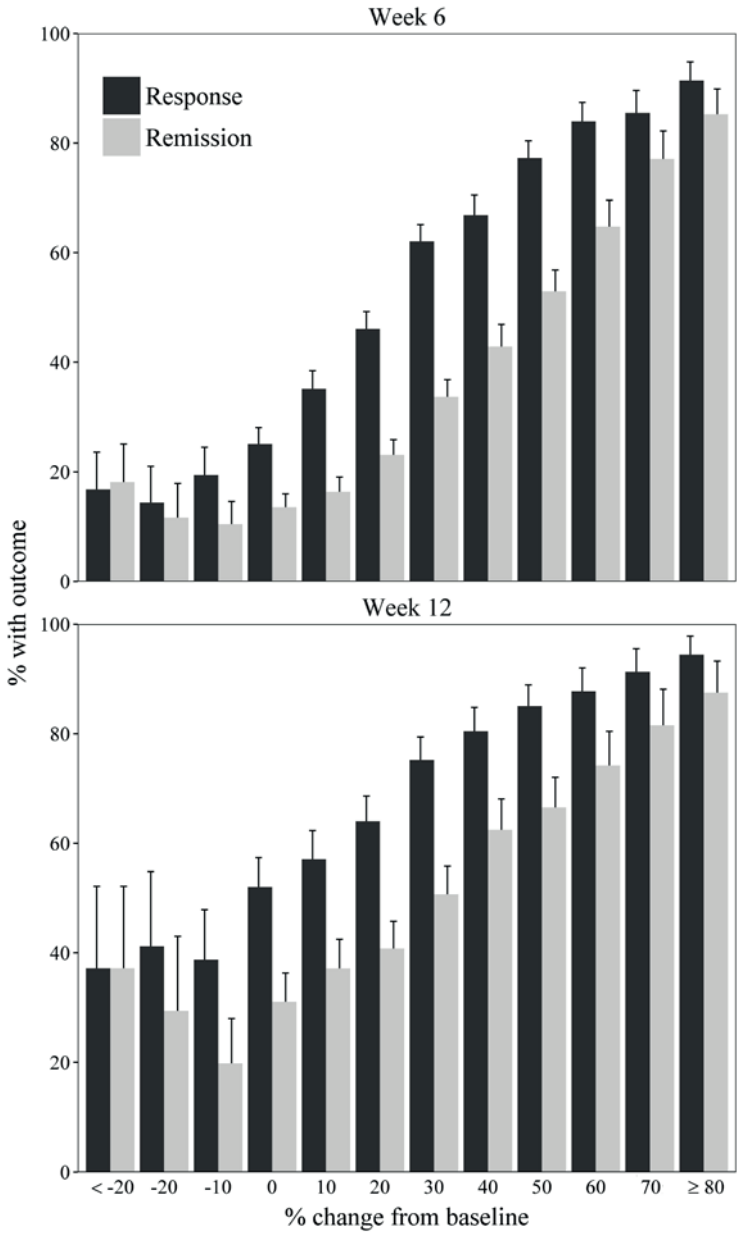

Figure 11.3: *Proportion of participants who responded or remitted according to the percentage improvement from baseline. Error bars indicate the confidence interval.*

predictor of good outcomes, regardless of treatment type.

While our results confirm the predictive value of early improvement, this value was still relatively limited, especially for longer-term outcomes. Some authors have suggested that non-improvers have virtually no chance of attaining remission and that these patients' treatment should be adapted [87], but our results indicate that these patients can still achieve good outcomes.

By 12 weeks of treatment, around 40% of patients who were predicted to have a poor outcome had achieved response and around 35% had achieved remission. These probabilities are comparable to those previously found in another large, 12-week trial (GENDEP) [95] and show that a patient's eventual outcome cannot be predicted with any certainty after two weeks of treatment. Indeed, one study found that the probability that non-responders would respond within the next two weeks was stable throughout the first twelve weeks of treatment, at around 15% [469].

A degree of caution in adapting treatment may therefore be warranted, all the more so because switching antidepressants is no more effective than continuing the same antidepressant [470]. A systematic review also found little evidence in favor of the effectiveness of early dose escalation, although escalation was clearly associated with reduced tolerability [471]. Other strategies, such as augmentation, may be more successful, but also result in decreased tolerability [472].

Such strategies might be appropriate for some patients without early improvement, for instance if a fast response is essential because of suicidality, but are likely to be premature for many patients. Given the limited predictive accuracy of models based on symptoms alone, inclusion of a broader set of predictors (e.g. psychiatric history, comorbidity, or adverse events) may be necessary to achieve better predictions.

Previous research has indicated that symptoms are not interchangeable and that the depression sum score could obscure important information [96]. Several studies have also found that early improvement in specific symptoms is associated with good outcomes [98, 99, 100, 463, 464, 465], in seeming contrast to our work. However, none of these studies included early improvement in the total score, so their findings are not directly comparable to ours. Furthermore, a variety of symptoms were associated with good outcomes, including general somatic symptoms, gastrointestinal symptoms, insomnia, depressed mood, agitation, loss of interest, feeling slowed down, and others, which also suggests that the association between early-improving symptoms and good outcomes is not particularly specific.

Our lasso regressions also tended to select most of the HAM-D items, rather than a few specific items, although some items were consistently not selected (suicidality and hypochondriasis). These results suggest that, with regard to early improvement, individual symptoms do not *add* meaningful predictive information to the sum score (especially when taken as continuous).

This may be because symptoms are actually more or less interchangeable in this regard and early improvement in any symptom is associated with good outcomes, or because symptoms are correlated and tend to improve together. However, it could also be related to the reliability of individual items. Single items are more strongly affected by random error than multi-item scales, for which the random error can balance out, which could degrade the predictive ability of a symptom. Furthermore, since our outcomes were derived from the HAM-D sum score, they are inherently dependent upon improvement in all individual items, although this would not, *a priori*, exclude differences in predictive ability, particularly if the probability or time course of improvement differs.

Our post-hoc analyses show that the association between early improvement and outcome is gradual. While a cut-off, such as $\geq 20\%$ improvement, may be easier to use in clinical practice, there is no major difference between patients on either side of this cut-off. The likelihood of response or remission does, however, seem to plateau as the percentage improvement by week 2 drops below around 10%. For instance, the likelihood of response by week 6 is only 17% for patients who deteriorate early in treatment, and the likelihood of remission is only 13%.

By week 12, however, around 39% of patients who deteriorate early in treatment have responded and 26% have remitted, which suggests that good outcomes are still possible for these patients (though less likely), if a longer period until remission can be tolerated. These results may therefore offer some guidance to clinicians who are faced with patients showing variable degrees of early improvement and need to decide between continuing, switching, or intensifying treatment.

Strengths and limitations

Among the strengths of our study is our large sample size, achieved through combining IPD. We used a rigorous approach to building predictive models, including using lasso to prevent over-fitting and using separate test data to examine model performance. We also examined multiple outcomes (response and remission) and both a short and a longer time frame (6 and 12 weeks).

A limitation of our study is that we did not take dosing schedules into account. One study has found that early improvement was more predictive when rapid, rather than slow, dose escalation was used [87]. However, there is only limited evidence for a dose-response relationship for second-generation antidepressants [381, 473, 474], and dose escalation usually also continues beyond two weeks in clinical practice.

We also took a complete-cases approach, since we were interested in predicting outcomes in patients who are receiving treatment. Our results therefore do not apply to participants who discontinue their medication and drop out of the trial. Because participants may discontinue due to lack of efficacy, the probability of response or remission may have been

overestimated somewhat. Another important reason for discontinuation was the presence of adverse events, which are also a highly relevant factor in weighing the (expected) risk-benefit ratio of treatment, but this was beyond the scope of this study.

An additional limitation is that our data were derived from clinical trials with strict inclusion and exclusion criteria. Hence, the study population represents only a subset of treatment-seeking patients, and participants may, on average, have better outcomes than patients seen in clinical practice [475]. Further research is therefore necessary to confirm that our results generalize to the broader patient population, including those with extensive comorbidity or chronic depression.

Finally, we chose the threshold of $\geq 50\%$ probability to assign participants to the response or remission category. This is a reasonable cut-off with the advantage of being independent of the data. However, a different cut-off could increase the negative predictive value (at the cost of positive predictive value). In principle, this might identify a group of participants with a lower probability of response or remission. However, because of decreasing specificity, this group would become progressively smaller as negative predictive value increases, which would limit clinical applicability.

In post-hoc analyses, we examined risk quintiles, which suggested that a small group of participants with poor outcomes at week 6 could be identified, but predictive accuracy was reduced for week 12 outcomes. Similar results were obtained while examining continuous early improvement, which suggests that this is the upper bound of predictive accuracy that can be achieved on the basis of symptoms alone.

Conclusions

Our results show that a model with only early improvement in the total score is about as predictive as models that also contain individual symptoms. Hence, clinicians need not focus on specific symptoms, but can gain as much information about the likelihood of a good outcome from improvement in the total score alone, particularly if improvement is interpreted as a continuous measure. However, the absence of early improvement does not rule out later response or remission with certainty. Therefore, adapting treatment because of limited improvement in the first two weeks would be premature for many patients.

Appendix

Table 11.3: *Supplemental table of studies*

		Duration	Dose	N (week 6)		Baseline score
Drug	Trial	(weeks)	(mg/day)	Placebo	Drug	Mean (SD)
Paroxetine	01/001 [476]	6	10 – 50	18	19	25.0 (3.2)
IR	02/001 - 004 [477]	6	10 – 50	100	112	23.5 (4.0)
	03/001 - 006 [273]	6	10 – 50	117	141	23.3 (3.7)
	7 [478]	6	10 – 60	7	8	25.1 (4.2)
	9 [479]	6	10, 20, 30, 40	31	262	22.5 (3.0)
	115 [480]	12	20	92	206	22.3 (3.6)
			Fluox 20		215	
	128 [481]	12	20	115	257	23.1 (3.8)
			Fluox 20		276	
	276 [482]	6	30	13	15	22.8 (3.9)
	279 [483]	6	30	7	14	20.8 (3.7)
Paroxetine	448 [189]	12	IR 20 – 50	94	173	23.3 (2.8)
CR			CR 25 – 62.5			
	449 [189]	12	IR 20 – 50	97	193	23.6 (3.1)
			CR 25 – 62.5			
	487 [484]	12	IR 10 – 40	97	189	22.1 (3.1)
			CR 12.5 – 50			
	810 [485]	8	12.5, 25	128	267	23.5 (3.1)
Duloxetine	HMAQ-A [486]	8	40 – 120	56	56	18.5 (4.4)
			Fluox 20		27	
	HMAQ-B [487]	8	40 – 120	60	67	18.1 (5.2)
			Fluox 20		29	
	HMAT-A [488]	8	40, 80	76	142	17.5 (5.3)
			Parox 20		73	
	HMAT-B [489]	8	40, 80	71	142	17.9 (5.2)
			Parox 20		66	
	HMAY-A [490]	8	80, 120	87	171	20.0 (3.7)
			Parox 20		79	
	HMAY-B [491]	8	80, 120	96	184	21.0 (3.6)
			Parox 20		89	
	HMBH-A [492]	9	60	102	99	21.2 (4.0)
	HMBH-B [493]	9	60	112	99	20.5 (3.4)
	HMBU [494]	12	60 - 120	-	137	23.1 (3.7)
			Ven 75 - 225		153	
	HMCQ [494]	12	60 - 120	-	133	22.3 (3.3)
			Ven 75 - 225		294	
	HMCR [495]	8	60	113	225	17.7 (5.0)
			Escit 10		237	
	HMCV [496]	8	60	-	183	21.2 (3.9)
			Parox 20		204	
	HMFA [497]	12	60	89	205	18.8 (6.3)

continued

Table 11.3: *Supplemental table of studies*

		Duration	Dose	N (week 6)		Baseline score
Drug	Trial	(weeks)	(mg/day)	Placebo	Drug	Mean (SD)
	HMFS-A [498]	8	60	103	221	22.9 (4.2)
	HMFS-B [498]	8	60	111	225	22.8 (4.7)
NCEs	NKD20006 [499]	8	Parox 20	95	83	24.5 (2.8)
	NKF100096 [500]	8	Parox 20 - 30	97	88	22.2 (5.6)

CR: controlled release; Escit: escitalopram (active comparator); Fluox: fluoxetine (active comparator); IR: immediate release; NCEs: new chemical entities; Parox: paroxetine (active comparator); Ven: venlafaxine (active comparator).

Table 11.4: *Model coefficients for response at week 6*

	Model			
Variable	Baseline	Total improvement	Item improvement	Item interactions
Intercept	-0.874 (0.457)	-0.952 (0.082)	-1.680 (0.100)	-1.271 (0.124)
Age	-0.059 (0.034)			
HAM-D item 1	0.736 (0.452)			
HAM-D item 6	0.135 (0.066)			
HAM-D item 9	0.052 (0.063)			
HAM-D item 12	0.058 (0.063)			
HAM-D item 16	0.107 (0.077)			
HAM-D item 17	0.157 (0.082)			
Total improvement		1.606 (0.067)	0.600 (0.094)	0.734 (0.099)
Improv item 1			0.351 (0.077)	0.326 (0.138)
Improv item 2			0.376 (0.069)	0.077 (0.146)
Improv item 4			0.182 (0.069)	-0.122 (0.112)
Improv item 5			0.302 (0.070)	0.219 (0.143)
Improv item 6			0.243 (0.070)	0.119 (0.093)
Improv item 7			0.197 (0.072)	-0.128 (0.125)
Improv item 8			0.148 (0.070)	-0.094 (0.117)
Improv item 9			0.220 (0.069)	0.172 (0.127)
Improv item 10			0.274 (0.070)	0.091 (0.149)
Improv item 11			0.231 (0.068)	-0.029 (0.134)
Improv item 12			0.195 (0.079)	0.066 (0.103)
Improv item 13			0.328 (0.073)	-0.044 (0.148)
Improv item 14			0.319 (0.081)	-0.013 (0.174)
Improv item 16				-0.120 (0.113)
Improv 1 × improv 2				0.034 (0.149)
Improv 1 × improv 5				0.014 (0.146)
Improv 1 × improv 9				-0.015 (0.144)
Improv 1 × improv 10				-0.025 (0.150)
Improv 1 × improv 14				0.163 (0.184)
Improv 2 × improv 8				0.220 (0.140)
Improv 2 × improv 10				0.075 (0.139)
Improv 2 × improv 11				0.202 (0.137)
Improv 2 × improv 13				0.173 (0.146)
Improv 4 × improv 7				0.333 (0.137)
Improv 5 × improv 9				0.130 (0.141)
Improv 5 × improv 10				-0.015 (0.140)
Improv 5 × improv 11				0.037 (0.141)
Improv 6 × improv 8				0.192 (0.142)
Improv 6 × improv 14				0.255 (0.171)
Improv 8 × improv 12				0.232 (0.163)
Improv 10 × improv 7				0.240 (0.142)
Improv 10 × improv 11				0.067 (0.138)
Improv 10 × improv 13				0.055 (0.149)

continued

Table 11.4: *Model coefficients for response at week 6*

	Model			
Variable	Baseline	Total improvement	Item improvement	Item interactions
Improv 11 × improv 4				0.268 (0.141)
Improv 13 × improv 7				0.142 (0.148)
Improv 13 × improv 16				0.788 (0.213)
Improv 14 × improv 13				0.217 (0.170)
Observations	4,847	4,847	4,847	4,847
Log Likelihood	-3,291.889	-2,982.251	-2,864.412	-2,834.873
AIC	6,601.778	5,970.503	5,760.824	5,749.746
BIC	6,660.153	5,989.961	5,864.602	6,009.191

AIC: Akaike Information Criterion; BIC: Bayes Information Criterion; HAM-D: Hamilton Depression Rating Scale.

Table 11.5: *Model coefficients for remission at week 6*

	Model			
Variable	Baseline	Total improvement	Item improvement	Item interactions
Intercept	-0.775 (0.071)	-1.880 (0.088)	-2.610 (0.229)	-2.170 (0.149)
Baseline score	-0.424 (0.037)	-0.493 (0.039)	-0.709 (0.047)	-0.732 (0.044)
HAM-D item 2			-0.167 (0.121)	
HAM-D item 12			0.160 (0.091)	
HAM-D item 13			-0.241 (0.165)	
HAM-D item 16			0.139 (0.166)	
Total improvement		1.591 (0.080)	0.395 (0.110)	0.692 (0.121)
Improv item 1			0.390 (0.088)	0.218 (0.170)
Improv item 2			0.539 (0.081)	0.107 (0.182)
Improv item 4			0.238 (0.073)	-0.127 (0.161)
Improv item 5			0.321 (0.074)	0.108 (0.171)
Improv item 6			0.205 (0.075)	-0.055 (0.170)
Improv item 7			0.283 (0.077)	-0.179 (0.181)
Improv item 8			0.274 (0.074)	-0.031 (0.142)
Improv item 9			0.192 (0.073)	-0.232 (0.149)
Improv item 10			0.274 (0.076)	-0.177 (0.161)
Improv item 11			0.251 (0.073)	0.046 (0.137)
Improv item 12			0.097 (0.099)	0.056 (0.107)
Improv item 13			0.442 (0.076)	0.008 (0.168)
Improv item 14			0.433 (0.081)	-0.129 (0.194)
Improv item 16			0.250 (0.181)	-0.007 (0.216)
Improv item 17			0.274 (0.125)	0.086 (0.140)
Improv 1 × improv 2				-0.197 (0.172)
Improv 1 × improv 4				0.147 (0.177)
Improv 1 × improv 5				0.106 (0.175)
Improv 1 × improv 6				-0.022 (0.175)
Improv 1 × improv 7				0.222 (0.173)
Improv 1 × improv 14				0.272 (0.203)
Improv 1 × improv 16				0.167 (0.232)
Improv 2 × improv 6				0.195 (0.149)
Improv 2 × improv 7				0.107 (0.158)
Improv 2 × improv 8				0.256 (0.150)
Improv 2 × improv 9				0.225 (0.150)
Improv 2 × improv 10				0.208 (0.153)
Improv 2 × improv 11				0.083 (0.147)
Improv 2 × improv 13				0.015 (0.154)
Improv 4 × improv 7				0.243 (0.153)
Improv 4 × improv 8				0.190 (0.150)
Improv 5 × improv 10				0.048 (0.151)
Improv 5 × improv 11				0.144 (0.145)
Improv 5 × improv 13				0.027 (0.149)
Improv 6 × improv 8				0.102 (0.149)

continued

Table 11.5: *Model coefficients for remission at week 6*

	Model			
Variable	Baseline	Total improvement	Item improvement	Item interactions
Improv 6 × improv 9				0.277 (0.147)
Improv 7 × improv 10				0.203 (0.155)
Improv 9 × improv 10				0.288 (0.154)
Improv 10 × improv 13				0.215 (0.156)
Improv 11 × improv 13				0.201 (0.147)
Improv 14 × improv 12				0.250 (0.184)
Improv 14 × improv 13				0.491 (0.167)
Improv 16 × improv 8				0.233 (0.203)
Improv 16 × improv 12				0.225 (0.210)
Improv 16 × improv 17				1.010 (0.334)
Observations	4,847	4,847	4,847	4,847
Log Likelihood	-2,947.784	-2,712.485	-2,547.361	-2,512.073
AIC	5,901.569	5,432.971	5,140.722	5,122.146
BIC	5,921.027	5,458.915	5,289.903	5,439.966

AIC: Akaike Information Criterion; BIC: Bayes Information Criterion; HAM-D: Hamilton Depression Rating Scale.

Table 11.6: *Model coefficients for response at week 12*

	Model			
Variable	Baseline	Total improvement	Item improvement	Item interactions
Intercept	0.773 (0.140)	0.081 (0.130)	0.007 (0.161)	-0.051 (0.155)
Age		-0.228 (0.065)	-0.235 (0.066)	-0.222 (0.066)
HAM-D item 4			-0.580 (0.134)	
Total improvement		1.202 (0.104)	0.394 (0.146)	0.452 (0.149)
Improv item 1			0.456 (0.126)	0.475 (0.163)
Improv item 2			0.306 (0.116)	0.010 (0.173)
Improv item 4			0.536 (0.129)	0.237 (0.201)
Improv item 7			0.168 (0.120)	0.119 (0.143)
Improv item 6				-0.416 (0.176)
Improv item 10			0.293 (0.116)	-0.331 (0.180)
Improv item 12				-0.285 (0.185)
Improv item 13			0.411 (0.124)	-0.046 (0.274)
Improv 1 × improv 4				-0.028 (0.239)
Improv 1 × improv 13				-0.042 (0.274)
Improv 2 × improv 10				0.505 (0.222)
Improv 4 × improv 13				0.195 (0.247)
Improv 10 × improv 6				0.675 (0.231)
Improv 10 × improv 12				0.793 (0.272)
Improv 13 × improv 2				0.138 (0.248)
Improv 13 × improv 6				0.494 (0.247)
Improv 13 × improv 7				0.161 (0.254)
Improv 13 × improv 10				0.127 (0.245)
Observations	1,961	1,961	1,961	1,961
Log Likelihood	-1,200.156	-1,125.716	-1,085.160	-1,076.193
AIC	2,404.312	2,259.431	2,192.319	2,196.386
BIC	2,415.474	2,281.756	2,253.712	2,319.172

AIC: Akaike Information Criterion; BIC: Bayes Information Criterion; HAM-D: Hamilton Depression Rating Scale.

Table 11.7: *Model coefficients for remission at week 12*

	Model			
Variable	Baseline	Total improvement	Item improvement	Item interactions
Intercept	0.305 (0.150)	-0.425 (0.160)	-1.034 (0.183)	-0.552 (0.226)
Age	-0.135 (0.059)	-0.142 (0.059)	-0.139 (0.061)	-0.065 (0.079)
Gender				-0.063 (0.149)
Baseline score	-0.270 (0.056)	-0.328 (0.058)	-0.398 (0.062)	-0.436 (0.066)
HAM-D item 3	-0.298 (0.102)	-0.231 (0.106)	-0.173 (0.109)	-0.160 (0.113)
HAM-D item 4	-0.322 (0.109)	-0.395 (0.113)	-0.588 (0.134)	-0.593 (0.138)
HAM-D item 16	0.355 (0.121)	0.361 (0.124)	0.430 (0.129)	0.474 (0.222)
Total improvement		1.151 (0.103)	0.096 (0.148)	0.261 (0.163)
Improv item 1			0.557 (0.125)	0.487 (0.207)
Improv item 2			0.388 (0.108)	-0.200 (0.246)
Improv item 4			0.480 (0.124)	0.194 (0.245)
Improv item 5			0.233 (0.108)	0.078 (0.151)
Improv item 6				-0.001 (0.190)
Improv item 7			0.322 (0.111)	-0.066 (0.263)
Improv item 8				-0.268 (0.158)
Improv item 9				-0.133 (0.191)
Improv item 10			0.244 (0.110)	-0.091 (0.170)
Improv item 11			0.167 (0.104)	-0.029 (0.175)
Improv item 13			0.268 (0.111)	-0.178 (0.213)
Improv item 14			0.262 (0.122)	-0.086 (0.258)
Improv item 15				0.010 (0.121)
Improv item 16				-0.294 (0.274)
Age × improv 8				-0.130 (0.107)
Age × improv 14				-0.128 (0.125)
Gender × improv 7				0.196 (0.213)
Improv 1 × improv 2				-0.015 (0.248)
Improv 1 × improv 4				-0.095 (0.241)
Improv 1 × improv 7				0.226 (0.246)
Improv 1 × improv 14				0.350 (0.294)
Improv 2 × improv 7				0.022 (0.229)
Improv 2 × improv 8				0.395 (0.216)
Improv 2 × improv 10				0.435 (0.218)
Improv 2 × improv 11				0.201 (0.210)
Improv 2 × improv 13				0.147 (0.228)
Improv 4 × improv 6				0.396 (0.217)
Improv 4 × improv 7				0.002 (0.220)
Improv 4 × improv 9				0.477 (0.215)
Improv 6 × improv 9				-0.788 (0.220)
Improv 7 × improv 9				0.451 (0.213)
Improv 10 × improv 13				0.289 (0.223)
Improv 11 × improv 5				0.251 (0.210)
Improv 13 × improv 6				0.445 (0.221)

continued

Table 11.7: *Model coefficients for remission at week 12*

	Model			
Variable	Baseline	Total improvement	Item improvement	Item interactions
Improv 15 × improv 16				0.631 (0.309)
Observations	1,961	1,961	1,961	1,961
Log Likelihood	-1,315.276	-1,249.787	-1,195.004	-1,164.499
AIC	2,644.551	2,515.574	2,424.008	2,416.998
BIC	2,683.620	2,560.224	2,518.889	2,662.571

AIC: Akaike Information Criterion; BIC: Bayes Information Criterion; HAM-D: Hamilton Depression Rating Scale.

Table 11.8: *Model performance for secondary analyses investigating interactions with treatment group*

		AUC			
		Week 6		Week 12	
Model	Interactions?	Response	Remission	Response	Remission
Baseline	No	0.61	0.66	0.63	0.62
	Yes	0.60	0.66	0.60	0.61
Total improvement	No	0.74	0.75	0.69	0.69
	Yes	0.74	0.75	0.68	0.68
Item improvement	No	0.78	0.79	0.72	0.72
	Yes	0.77	0.78	0.70	0.70
Item interactions	No	0.78	0.79	0.72	0.73
	Yes	0.77	0.78	0.69	0.72

AUC: Area under the (receiver operating characteric) curve.

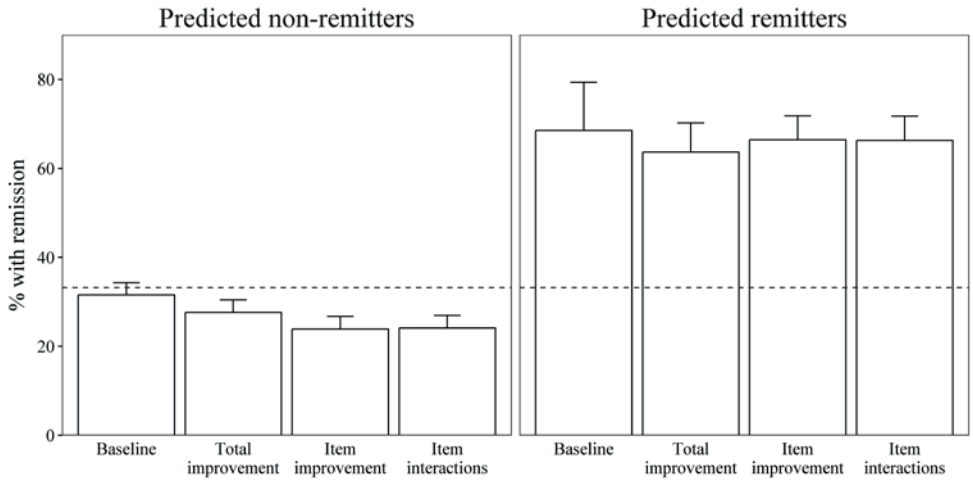

Figure 11.4: *Actual probability of remission at week 6 according to participants' predicted outcome (non-remission vs. remission). The dashed line indicates the baseline probability of remission. The models predicted non-remission for 96% (baseline), 85% (total improvement), 78% (item improvement), and 78% (item interactions) of participants.*

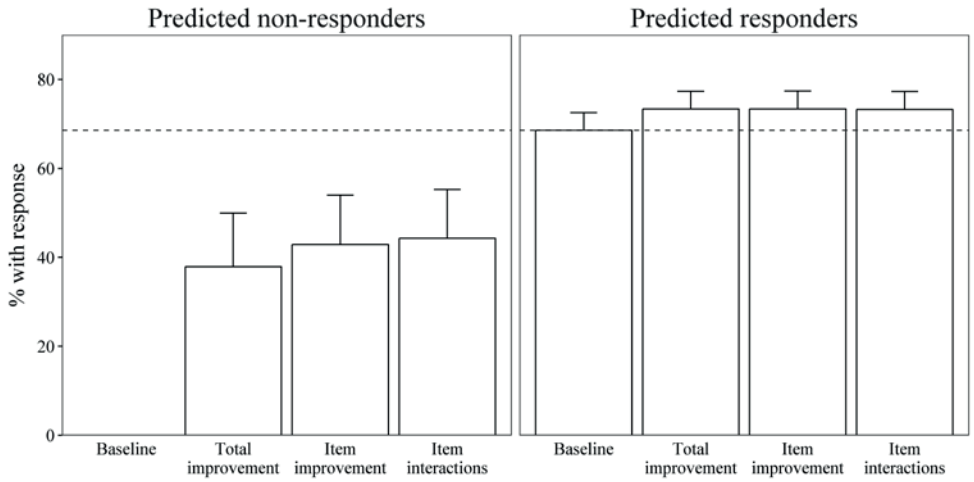

Figure 11.5: *Actual probability of response at week 12 according to participants' predicted outcome (non-response vs. response). The dashed line indicates the baseline probability of response. The models predicted non-response for 0% (baseline), 13% (total improvement), 16% (item improvement), and 16% (item interactions) of participants.*

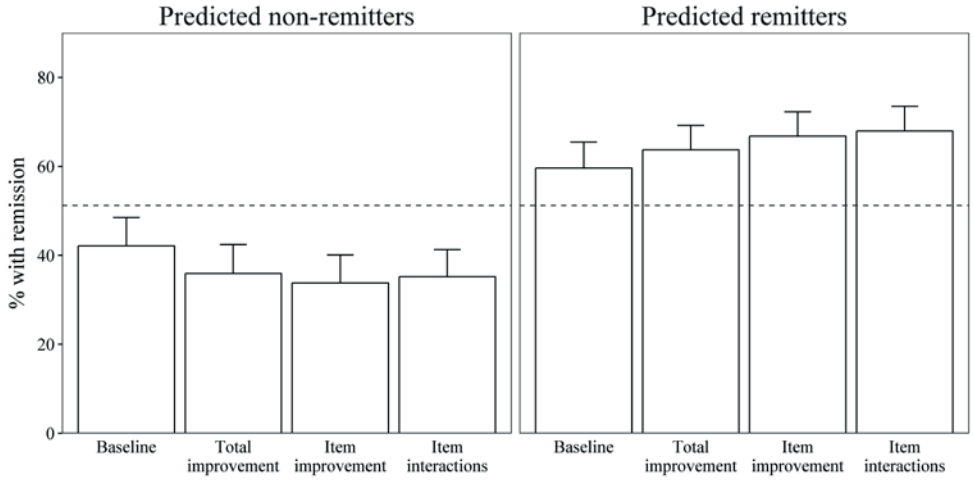

Figure 11.6: *Actual probability of remission at week 12 according to participants' predicted outcome (non-remission vs. remission). The dashed line indicates the baseline probability of remission. The models predicted non-remission for 48% (baseline), 45% (total improvement), 47% (item improvement), and 51% (item interactions) of participants.*

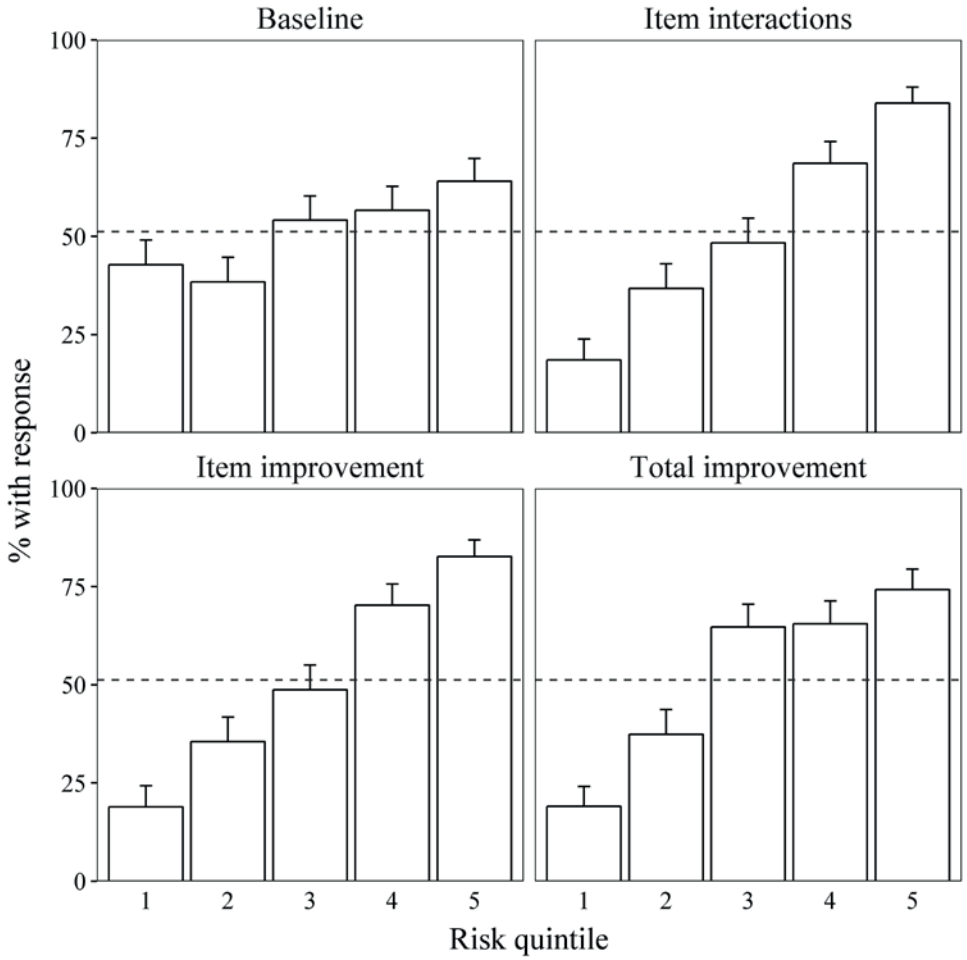

Figure 11.7: *Actual probability of response at week 6 according to risk quantiles and model. For each model, participants' predicted probability of response was used to divide participants into quintiles of "risk", with the lowest quintile having the lowest predicted probability of response.*

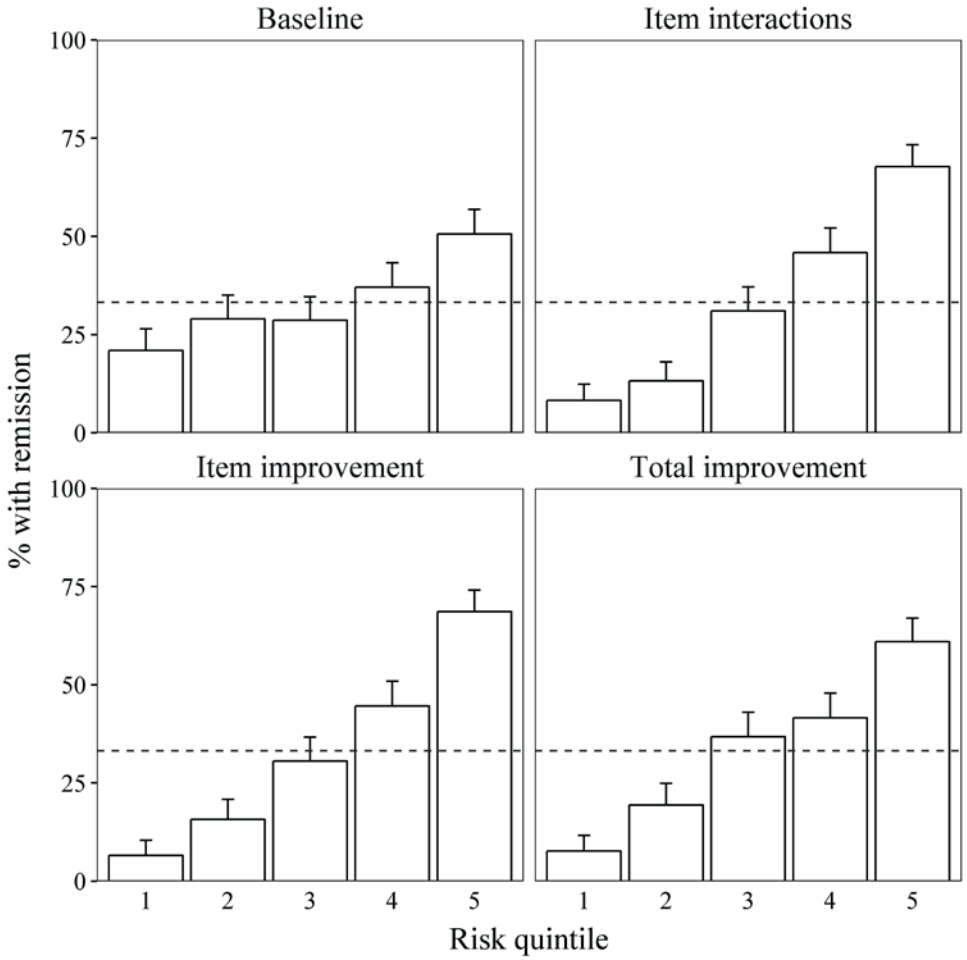

Figure 11.8: *Actual probability of remission at week 6 according to risk quantiles and model. For each model, participants' predicted probability of response was used to divide participants into quintiles of "risk", with the lowest quintile having the lowest predicted probability of response.*

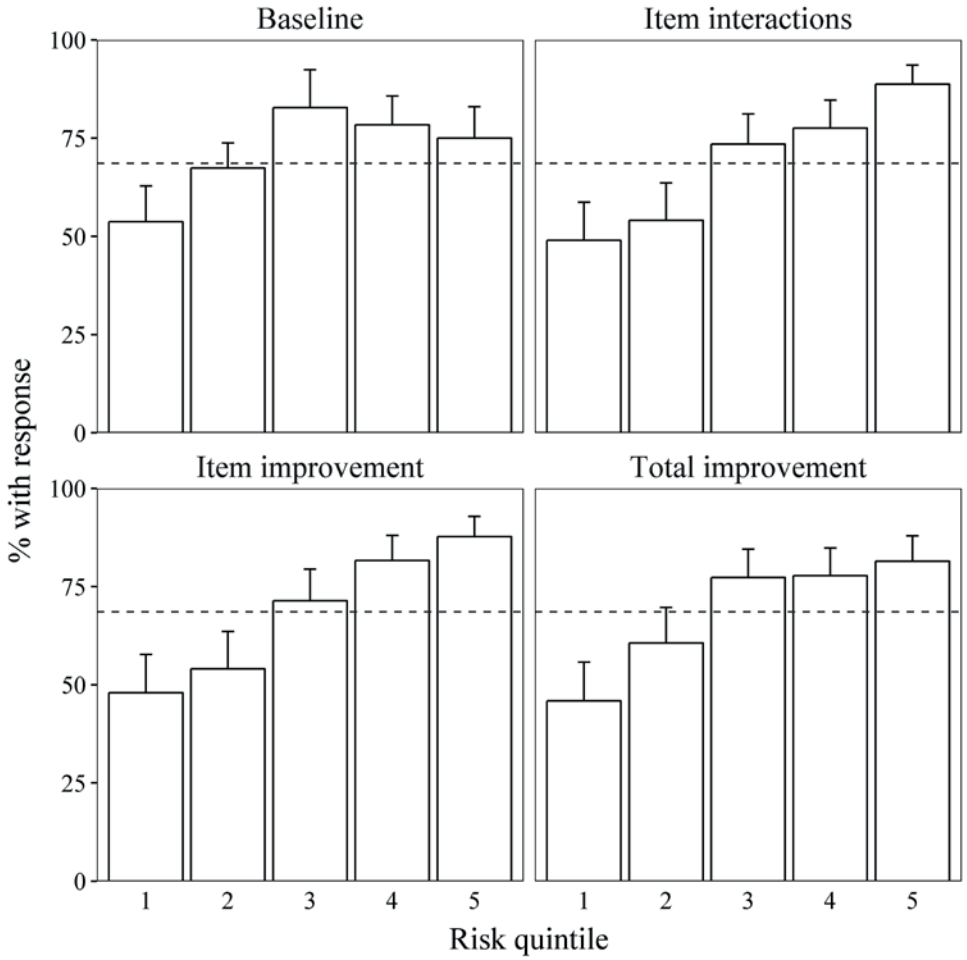

Figure 11.9: *Actual probability of response at week 12 according to risk quantiles and model. For each model, participants' predicted probability of response was used to divide participants into quintiles of "risk", with the lowest quintile having the lowest predicted probability of response.*

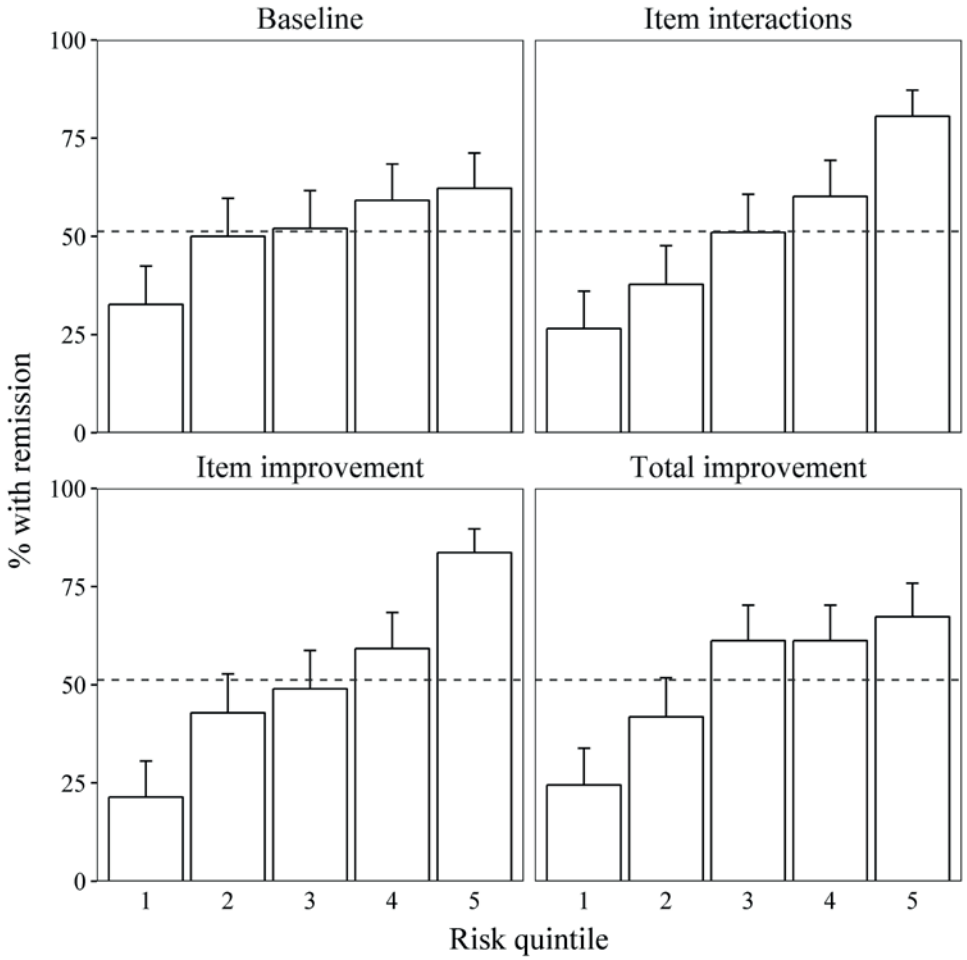

Figure 11.10: *Actual probability of remission at week 12 according to risk quantiles and model. For each model, participants' predicted probability of response was used to divide participants into quintiles of "risk", with the lowest quintile having the lowest predicted probability of response.*

Chapter 12

General discussion

In this thesis, I have re-examined the evidence for the treatment of depression and anxiety. By doing so, I aimed to answer questions about these treatments that have remained unanswered so far, in spite of the hundreds of trials that have been conducted over the past decades. My focus was, firstly, on examining the impact of reporting and citation biases on the literature, and secondly, on investigating clinical predictors of treatment response.

In this discussion, I will first briefly summarize the main findings of the different chapters. Then I will place my findings on reporting and citation bias in a broader perspective, discuss possible solutions to these issues and their effectiveness (insofar as these solutions have already been implemented), and consider further work that is necessary. Subsequently, I will focus on treatment efficacy, the clinical relevance of the true efficacy and safety of antidepressants, and what kind of research is needed now to improve outcomes for depressed and anxious patients.

Summary of main findings

In the first part of this thesis, I examined the presence of reporting and citation biases in the literature on the treatment of depression and anxiety. For several of these chapters, I used Food and Drug Administration (FDA) drug application packages. Because pharmaceutical companies are required to preregister these trials with the FDA and must submit their results, regardless of trial outcome, these packages contain a complete and unbiased overview of all pre-marketing trials conducted in the pursuit of approval for a specific drug for a specific indication.

Building upon previous research in depression, chapter 2 showed that study reporting bias, outcome reporting bias, and spin are present in the literature on antidepressants for the short-term treatment of anxiety disorders. According to the FDA, 72% of these antidepressant trials were positive, but 96% of published articles were positive. Of the negative trials, seven remained unpublished, three were published with outcome reporting bias, and three others were published with spin. Only three trials were clearly published as negative. As a consequence of these biases, the effect size of antidepressants for anxiety disorders had been overestimated by 15%, although this difference was not statistically significant.

Chapter 3 examined reporting bias for harm outcomes in antidepressant trials for anxiety as well as depression. No bias in the reporting of discontinuation rates was found. However, nearly two-thirds of journal articles did not mention serious adverse events (SAEs) at all. Of the articles that did mention SAEs, the majority contained discrepancies with the FDA or did not include any descriptions of the SAEs, including one article that failed to mention two suicides in the drug group. Together, these findings show that the published literature is an unreliable source of information, both regarding the efficacy of

antidepressants and regarding their safety (especially where SAEs are concerned).

In chapter 4, I investigated the phenomenon of pooling otherwise unpublished trials for publication. These pooled-trials publications were very common, but very few (12%) of these publications had as their primary aim to present data on the included trials' primary research question (comparing the efficacy of the drug to placebo), and even fewer (3%) presented efficacy data for the primary research question for individual trials. Even though the vast majority of these pooled-trials publications included one or more negative trials, only 3% had a negative conclusion. Consequently, pooled-trials publications flood the literature with positive results for secondary questions while obscuring the negative findings for the primary outcome.

In chapters 5 and 6 I examined spin (or positive focus) and citation bias in the literature on 5-HTTLPR. Within both the amygdala activation and the gene-environment interactions literature, articles with negative findings often presented positive conclusions in their abstract. Both truly positive articles and positively presented articles receive more citations than articles with negative findings and negative conclusions. This effect was stronger within the amygdala activation literature than the gene-environment literature, perhaps because of the greater controversy surrounding gene-environment interactions. These results show how negative findings, even when published, can remain relatively invisible, which contributes to a more positive view of the evidence base for an effect than is warranted.

Chapter 7 investigated the cumulative effect of reporting and citation biases on the evidence base for psychotherapy and antidepressants for depression. While study publication bias is a well-known phenomenon by now, this chapter demonstrates the pernicious cumulative effect of study publication bias, outcome reporting bias, spin, and citation bias. Starting with a cohort of 105 antidepressant trials, of which 52 were considered negative by the FDA, I finally found that only four articles unambiguously report that the antidepressant was not effective. Positive trials were also cited three times as frequently as negative trials (92 versus 32 citations on average). Each of these biases makes it more difficult to discover negative results, and together, they can render the vast majority of negative results virtually invisible.

In chapter 8, I studied whether the evidence is actually put into practice. Guideline committees critically appraise and synthesize all relevant evidence in order to arrive at treatment recommendations. These guidelines can be seen as representing the "state of the art" for any given topic. However, adherence to the guidelines for antidepressant initiation in children and adolescents was very poor. Physicians preferred citalopram, an antidepressant that has never been shown to be effective in these age groups, over the recommended fluoxetine, which has the best available evidence in favor of its efficacy and safety in young people. Starting doses were also higher than recommended, especially in teens, who were usually prescribed adult starting doses. These findings show that translation of the evidence base to clinical practice often fails, thus undermining the

goals of evidence-based medicine.

The second part of this thesis aimed to investigate who benefits from antidepressants. In chapters 9 and 10 I examined initial severity as a predictor of antidepressant response compared to placebo in anxiety disorders. In chapter 9, I used aggregate data from 56 clinical trials for GAD, SAD, OCD, PTSD, and panic disorder and found no evidence that greater initial severity was associated with a better response to antidepressants. However, because the use of trial-level data to investigate a patient-level characteristic, like initial severity, can be subject to the ecological fallacy, and because reduced power means that interactions may have been missed, I used individual participant data (IPD) to investigate this question in chapter 10. Here, I found that initial severity was associated with antidepressant response for GAD and panic disorder, but not for SAD, OCD, or PTSD. These findings suggest that the benefit of antidepressants is relatively small at low severity for GAD and panic disorder and the risk-benefit ratio may therefore be unfavorable for patients with mild GAD or mild panic disorder.

In chapter 11, I examined whether early improvement in individual depressive symptoms could enhance the predictive power of a model already containing early improvement in the total score. There was limited evidence that this was the case, which suggests that clinicians can gain about as much information about a patient's likelihood of responding or remitting to antidepressants from the total score. However, the predictive utility of this model was still relatively limited, particularly for outcomes after twelve weeks (rather than six weeks) of treatment, which suggests that we cannot predict a patient's likelihood of a good response with much certainty by two weeks of treatment.

Evidence-b(i)ased psychiatry[1]

In the first part of my thesis, I showed that reporting and citation biases are prevalent within the antidepressant and psychotherapy literature. These biases complicate assessment of the risks and benefits of these treatment options. While some biases, like study publication bias and outcome reporting bias, can and do affect the results of systematic reviews and meta-analyses [325], spin and citation bias, in principle, do not. However, the interpretation of meta-analytic results is not perfectly straightforward and objective: these interpretations in fact vary widely, perhaps dependent on readers' prior beliefs [501, 502]. Hence, meta-analyses are not a foolproof remedy against the effects of spin and citation bias.

The problem of study publication bias was first recognized over fifty years ago [36]. It is therefore disconcerting to realize how little progress appears to have been made in the decades following its first identification. Only recently have medical journals, funders, and

[1] With acknowledgment to Melander and colleagues [27], who coined the term "evidence-b(i)ased medicine".

governments begun to take measures to combat study publication and outcome reporting bias. The primary weapon in their arsenal is mandatory pre-registration of clinical trials. In 2004, the International Committee of Medical Journal Editors (ICMJE) made pre-registration a requirement for publication in ICMJE journals, an important first step [330]. However, many medical journals do not adhere to the ICMJE guidelines, and many others also accept retrospective registration [332, 503]. Retrospective registration, particularly when it takes place after ascertainment of the primary outcome, is virtually useless when it comes to preventing study publication and outcome reporting bias.

In 2007, the FDA Amendments Act (FDAAA) was signed into law. This Act requires prospective registration of "applicable clinical trials" and empowers the FDA to levy fines against non-compliant investigators [504]. Although the FDA has never yet exercised its power to penalize investigators for failure to register, this regulatory requirement may nevertheless prove to be more effective than the ICMJE requirement, since it applies to all applicable clinical trials, regardless of the publishing aspirations of the investigators, and has the potential force of the law behind it.

Importantly, the FDAAA also requires investigators to post a public summary of results within one year of completion of the trial, which would ensure that trial results are easily available even if the trial is never submitted for publication. However, compliance with this requirement is dramatically low: only 13% of all trials report results within the 12-month deadline and only 38% report them at any time [505]. Ironically, given the particular attention paid to selective publication by pharmaceutical industry, industry-sponsored trials are more likely to report results on time than academic trials, perhaps because pharmaceutical companies have more resources to bring to bear to meet this requirement.

Compliance may be low in part because penalties have never yet been enforced and in part because the requirement for results reporting conflicts with the requirement by many journals that results have not yet been disseminated (although many journals have since clarified their requirements to permit basic results reporting on sites like ClinicalTrials.gov) [506].

An unfortunate weakness of the FDAAA is that it only includes trials of drugs or medical devices, thus excluding trials of behavioral interventions like psychotherapy. Since bias is equally problematic in these trials [35], this is a regrettable omission. An additional, though legally inevitable, problem is that the FDA only has jurisdiction over trials with at least one center located in the United States or that are performed in the service of an application for marketing approval in the United States.

Mandatory requirements for prospective registration have the potential to eliminate publication bias and outcome reporting bias from the medical literature, but they require "constant vigilance" in order to succeed. Clinical trial registries must be examined thoroughly to detect trials that have been registered but not published. While particular

attention should be paid to trials that do not even report basic results, we should not be satisfied with trials remaining unpublished (and essentially invisible) otherwise. Basic results reporting on ClinicalTrials.gov is important, but it cannot be considered equivalent to publishing a trial in a high-impact medical journal that will be read by thousands of researchers and clinicians. Furthermore, unless peer reviewers and editors carefully scrutinize pre-registrations during the review process, it is likely that outcome switching will continue to occur.

The medical literature could also benefit from importing the concept of registered reports [507], an idea that originates within the field of psychology and that is more of a "carrots" (reward) approach compared to the "sticks" (punishment) of the FDAAA and the ICMJE policy.

Registered reports take the logical next step beyond pre-registration. A researcher designs a study, writes the introduction and methods section of the article that will eventually result from the study, and submits this to the desired journal. The study undergoes peer review and, if the research question is considered interesting and the design suitable, the study receives "in-principle acceptance", that is: if the study is conducted as proposed, it will be published by the journal regardless of the results. The researcher only performs the study after in-principle acceptance has been obtained.

Registered reports, hence, are a form of "results-free" reviewing, which ensures that papers are not rejected because of negative results. This is an aspect that is still missing in a system of prospective registration, which ensures that we are aware of all trials that have been conducted (and potentially of their results, if compliance with mandatory results reporting improves) but does nothing to prevent negative results from being published later than positive results (time-lag bias), in lower-impact journals than positive results (place of publication bias), in languages other than English (language bias) [46], and so on. It cannot even prevent non-publication altogether (although perhaps nothing, short of extremely strict enforcement of high penalties, is 100% guaranteed to prevent study publication bias).

Another benefit to registered reports is that it invites peer review of the study design before the study has been conducted, at a stage when fatal or minor flaws can still be rectified. The current system of peer review after completion of the study, on the other hand, is more reminiscent of the famous quote by Ronald Fisher [508]: "To consult the statistician after an experiment is finished is often merely to ask him to perform a post mortem examination. He can perhaps say what the experiment died of." The registered report format is difficult to adapt to all possible study types (e.g. secondary analyses of pre-existing epidemiological cohorts, exploratory studies), which may limit its implementation in some fields, but it is perfectly suited to clinical trials.

There is some reason to be cautiously optimistic about the future of the medical literature. Comparing the trials for newer antidepressants to those of older antidepressants (chapter

7), for example, we found that the newer, negative trials were much more likely to be published than older negative trials. This may, to some extent, be related to the extremely critical and focused attention that has been paid to biased reporting of antidepressant trials in particular, but other studies have also found an increase in publication rates in recent years [509].

If results reporting on sites like ClinicalTrials.gov is included, disclosure rates for industry-sponsored trials were >90% in recent years [510]. It is possible, therefore, that the problems that motivated our work in chapter 2, investigating study publication bias and outcome reporting bias in antidepressant trials for anxiety disorders, will largely disappear in the coming years, which would be a major step forward for evidence-based psychiatry.

Rectifying the existing literature, however, has proven to be a difficult task. None of the articles that we identified as biased have since been retracted or corrected [511]. Even the article reporting the results of Study 329, the trial of paroxetine for adolescent depression that has been embroiled in controversy ever since its first publication and that has been the focus of a dedicated campaign to correct the record, has neither been retracted nor corrected [512].

It is also unlikely that universal prospective registration will effectively combat all biases. In particular, it is unlikely to ameliorate the problems of spin and citation bias at all, and it is possible and perhaps likely that financial incentives to suppress negative results will lead pharmaceutical sponsors to seek out other methods, such as publishing negative results in journal articles that are not widely read or not even indexed in databases like PubMed or EMBASE. One of the negative trials in our analysis in chapter 7, for instance, was published in the Journal of Drug Assessment, which was not indexed in PubMed until very recently. We only discovered this publication because it happened to be mentioned in a review [513].

Although there is a widespread sense that negative results are more difficult to publish, and it might be argued that this is why this trial was published in a journal where it was virtually guaranteed to go unnoticed, this argument does not seem very plausible in the age of PLOS ONE, BMJ Open, and comparable journals, which publish research regardless of its clinical relevance and newsworthiness.

Conversely, pharmaceutical companies make an effort to publish positive results from their studies, by pooling trials, slicing and dicing the results, and publishing highly redundant articles that appear to serve primarily to keep a drug in the spotlight, as shown in chapter 4, in which we investigated pooled-trials publications. The added scientific value of these publications often seems limited, and instead these publications often seem to represent "minimal publishable units". For duloxetine, for instance, there are separate papers investigating the efficacy of duloxetine in African-American compared to Caucasian patients, and in Hispanic compared to Caucasian patients [174]. Both papers conclude that duloxetine works equally well regardless of race or ethnicity, and we can

wonder whether they would have been published if they had reached any other conclusion.

The medical literature, in this case, has been co-opted by pharmaceutical interests. However, medical journals are not just being duped, but are also complicit in these practices and have their own (financial) conflicts of interest, both directly (e.g. selling of reprints to pharmaceutical companies) and indirectly (e.g. maintaining or achieving a high impact factor through publishing highly citable papers) [514].

Biased and incomplete reporting of harm outcomes, which we investigated in chapter 4, also remains a problem that is less likely to be solved by universal prospective registration. To some extent, poor reporting of harms may be because researchers and clinicians are simply less interested in these outcomes. Most trials are powered to detect a significant difference in the primary outcome and are underpowered to study (rare) harm outcomes, which means the trials are not very informative in this regard. Nevertheless, full reporting is essential in order to enable meta-analyses, which can compensate for the small sample sizes of individual trials.

Word count limits may have been a good reason for the limited reporting of the many and diverse harm outcomes in the past, since tables of common adverse events alone may take up several pages. However, as most journals are now primarily electronic, authors can include the full information in the online supplemental information. With regard to serious adverse events, one important take-away from our study in chapter 4 is that causal attribution should play no role in the reporting of these events. Even events that are thought to be completely unrelated to the drug should be reported, since it is impossible to definitively establish causality in individual cases.

As important as the biased reporting of SAEs, however, is the complete absence of reporting on these events, which occurred in the majority of journal articles. Since regulatory authorities require investigators to monitor SAEs in drug trials, journals can be sure that these events were recorded and should require authors to report on the number and nature of SAEs.

Some initiatives, like accreditation or ratings, could encourage pharmaceutical companies to grant their negative results as much visibility as their positive results [515], perhaps following the example of the Access to Medicine Index [516]. This Index charts to what extent the largest pharmaceutical companies are working to make medication and vaccines more accessible to people in low- and middle-income countries. In this way, it makes visible which companies are doing well and which companies are doing poorly, and incentivizes companies to do better. A "Good Scientific Practice" index might similarly provide an incentive for companies to adhere to scientific best practices: to publish all studies, regardless of results; to stick to the statistical analysis plan; and to interpret results objectively and fairly, without spin. Like the Access to Medicine Index, such an index could be funded by non-profit foundations and governments.

Much more radical propositions, however, have also been put forward. In particular, there

is a case to be made for taking clinical trial programs out of the hands of pharmaceutical companies and placing them into the care of governmental institutions or universities [517]. The current situation is rife with conflicts of interest: pharmaceutical companies, who stand to benefit from a drug if and only if it is approved by the authorities and prescribed by physicians, carry the responsibility for conducting the trials that are meant to show that the drug is safe and effective. Strict regulation can mitigate this inherent conflict of interest, but complete elimination of pharmaceutical company sponsorship, though potentially expensive, might be the only comprehensive solution. This is, however, unlikely to happen, especially in this era of "public-private partnerships" [517].

Elimination of pharmaceutical sponsorship also does not solve the conflicts of interest of non-industry researchers. Academic researchers do not usually have the same financial incentives to suppress negative results as pharmaceutical companies do, although some do (e.g. developers of a psychotherapy approach who earn royalties from sales of the treatment manual). They may, however, have non-financial conflicts of interest, for instance due to developing the treatment in question, a concept known as "allegiance bias" in psychotherapy research [518].

Academic researchers also have clear career incentives to publish frequently and in high-impact journals [519]. These high-impact journals, however, are most interested in novel, surprising, counter-intuitive, and (usually) positive results, one possible reason why the retraction rate is higher in higher-impact journals [520]. Unfortunately, study results are the one aspect of a study that a researcher has no control over – at least not unless he or she engages in outcome switching, so-called p-hacking, or other questionable research practices. Career incentives to get published often and in "good" journals, therefore, are often misaligned with good science [321].

Even in the absence of any questionable research practices, this process can lead to the "natural selection of bad science", as researchers whose methods are less rigorous (e.g. smaller sample sizes) do better career-wise, while science itself suffers [521, 522]. However, it is at least in principle possible to re-align these career incentives to facilitate instead of hinder good science; this is not the case for industry research, which will always primarily serve commercial, rather than scientific interests.

Instead of rewarding scientists for results and emphasizing simple quantitative measures of success that can easily be "gamed", like the h-index [521], scientists must be rewarded for methodologically sound and clinically relevant research, regardless of the results [523]. As a first step, this may require greater openness with regard to data, methods, analysis code, and so on, in order to enhance reproducibility and other researchers' ability to appraise the work.

While methodological soundness is, to some extent, a subjective term, most research areas do have norms for what is considered adequate, which are sometimes formalized (e.g. in the form of reporting guidelines) and sometimes remain implicit. As argued

by Moore and colleagues, it is difficult, if not impossible, to judge whether research is "excellent" as opposed to merely adequate, but not as difficult to tell apart "sound" and "unsound" research [523]. Importantly, methodologically sound research should be informative (e.g. sufficiently powerful) regardless of whether the results reach the arbitrary threshold of "$p < 0.05$".

When negative results are no longer looked down upon, the need to spin results may also lessen. A re-orientation toward methodological soundness might also decrease the tendency to preferentially cite positive findings, although it seems likely that citation bias will be one of the more difficult biases to combat, since it is not amenable to relatively straightforward solutions like pre-registration and since it is probably human nature to prefer positive (and potentially actionable) findings. Chapters 5, 6, and 7 showed that both spin and citation bias are currently highly prevalent in the literature, both within observational research on the etiology of depression (chapters 5 and 6) and within the clinical trial literature (chapter 7).

The presence of spin, in particular, exemplifies the failure of peer review, since this problem is, in principle, easy enough to detect (especially in clinical trials), perhaps even easier than outcome reporting bias since it does not even require the reviewer to look up the clinical trial registration. Peer review, however, often fails [524]. Peer reviewers could, in theory, also be tasked with mitigating citation bias, but it is likely that peer reviewers are not aware of all the relevant literature themselves, so this is not a very feasible demand, except perhaps in very small research fields.

In many cases, the most effective defense against spin and citation bias might still be the performance of high-quality, unbiased meta-analyses, although this is not an actual solution but rather a harm-reduction method. This is also likely to be a more effective approach in fields where the evidence base is clear, homogeneous, and uncontroversial. In the literature on the 5-HTTLPR and stress interaction, for instance, several meta-analyses have been performed, reaching opposite conclusions [305, 306, 307, 308], and these meta-analyses only seem to have polarized the field further, rather than leading to consensus.

In this case, where meta-analyses conflict, researchers' prior beliefs probably play a large role in determining which meta-analysis they place their faith in. A recently published, very large, well-conducted collaborative meta-analysis (with a total sample size of 38,802) that found no evidence for an interaction between 5-HTTLPR and stress in the development of depression might finally put the debate to rest, however [525]. Even in cases where meta-analyses effectively establish a consensus, though, citation bias remains problematic to the extent that it incentivizes researchers, who are evaluated on the basis of citation metrics, to produce positive findings.

While the evidence base is often tainted by biases, it remains the best resource that we have to guide clinical decision-making. In chapter 8, however, I found that adherence

to evidence-based guidelines for antidepressant initiation in young people was very poor. Clinical practice is often resistant to change [365], but a major change in physicians' habits for antidepressant initiation in young people did take place in the first decade of the 21st century, since physicians abandoned paroxetine after it became embroiled in controversy and negative media attention due to treatment-emergent suicidality. However, they started prescribing citalopram instead, not fluoxetine. It is possible that the guideline, which was published in 2009, appeared too late to take full advantage of this period of shifting prescription habits to spur physicians on to switch to fluoxetine.

Another problem, specific to child and adolescent psychiatry, is that the vast majority of research has been performed in adults and there is little evidence specifically for pediatric patients, which may lead physicians to extrapolate from the (relatively solid) evidence in adults instead of working from the (relatively shaky) evidence in children and adolescents. However, such extrapolation is risky since children are not little adults. In the case of antidepressants, this is especially clear, as treatment-emergent suicidality appears to be an issue primarily in young people [190, 191].

This study, looking at antidepressant initiation in young people, is not an isolated example of poor adherence to guidelines [526], which shows that transfer of the evidence into practice is far from guaranteed. Many factors can affect the likelihood of guideline adherence, some of which are controllable by guideline developers (e.g. clarity of the recommendation), while others are not (e.g. a stable and uncontroversial evidence base) [361, 527].

It is clear, however, that guidelines must be actively disseminated and implemented, not just developed [365]. No matter how good the evidence base or how many measures are taken to combat bias, it is only when the evidence is actually put into practice that the ideals of evidence-based medicine are achieved.

The effectiveness of treatment

When the first selective serotonin reuptake inhibitors (SSRIs), such as fluoxetine and sertraline, were developed in the late eighties and early nineties, expectations were high: these antidepressants were received as wonder drugs [528]. In the quintessential book from this period, Listening to Prozac, Peter Kramer, a psychiatrist, related the stories of patients who did not just stop being depressed when they started taking fluoxetine (Prozac), but who became "better than well" [529]. Kramer speculated that these drugs might herald an era of "cosmetic psychopharmacology", in which healthy people would take psychotropic drugs to improve their personalities.

To some extent, Kramer was right: fully 12% of adult Americans took antidepressants in 2013 [24]. Among middle-aged women (40 – 59 years old), 23% took antidepressants in the period 2005 – 2008 [530], a percentage that has likely increased even further in the

decade since. Although depressive and anxiety disorders are relatively common in women, it is difficult to believe that a quarter of middle-aged women suffer from sufficiently severe and persistent depression or anxiety to warrant antidepressant treatment.

The SSRIs' image as wonder drugs, however, has also become tarnished, and the pendulum has swung the other way, with some now arguing that these drugs do not work at all or even make things worse [531, 532, 533]. The blame for the backlash against antidepressants can be placed, at least in part, on the pharmaceutical industry, since their practice of hiding unfavorable results enabled the overhyping of antidepressants and the subsequent letdown when both antidepressant efficacy and safety were revealed to be less impressive than the marketing had suggested.

However, antidepressant efficacy is unlikely to be zero. Our best estimate for the (unbiased) effect size of antidepressants compared to placebo for the acute treatment of depression and anxiety is a standardized mean difference (SMD) of around 0.30 – 0.35, based upon the meta-analysis by Turner and colleagues [19] and our own meta-analysis, included in this thesis (Chapter 2). This effect size might still have been overestimated slightly as a consequence of unblinding: some antidepressant-treated participants are likely to deduce that they are taking the active drug because they experience side effects. On the other hand, placebo-treated participants are also rather likely to experience adverse events that align with the anticipated side-effect profile of the drug, perhaps as a consequence of a "nocebo" effect. Placebo-treated participants in trials of tricyclic antidepressants (TCAs), for instance, experience more "typical" TCA side effects, such as vision problems and dry mouth, than placebo-treated participants in SSRI trials [534].

While some have suggested that the effect of antidepressants could be wholly explained by unblinding [535], this seems improbable, for a few reasons. First, some putative antidepressants do, in fact, fail in the course of development. Despite having side effects, these drugs cannot be shown to have antidepressant or anti-anxiety activity. Some are subsequently repurposed for other disorders (e.g. atomoxetine, which is now approved and used for attention-deficit hyperactivity disorder (ADHD) [536]; or flibanserin, a controversial drug recently approved by the FDA for hypoactive sexual desire disorder [537, 538]). Secondly, at least one approved antidepressant, reboxetine, has been shown to be less effective than SSRIs and is not even statistically significantly more effective than placebo, despite having worse tolerability than SSRIs [101].

Finally, although early work found that experiencing side effects was associated with better outcomes (suggesting that unblinding might play a role) [539], more recent studies have not confirmed any association between the proportion of patients with adverse events and trial outcome [540, 541]. Most recently, an IPD analysis found antidepressants to be effective both in patients with and without adverse events, and also found no association between adverse event severity and antidepressant efficacy, providing strong evidence against the hypothesis that unblinding explains the effect of antidepressants [542].

An effect size of around 0.3 cannot be considered large, but it is consistent with effect sizes found in other areas of psychiatry and general medicine [421]. The corresponding number needed to treat (NNT) is around 8 to 10, that is, out of every 8 to 10 patients treated with an antidepressant, one will have a better outcome than they would have had with placebo treatment.

It is, however, important to keep in mind that treatments do not actually have effect sizes; they only have effect sizes relative to a control condition. Consequently, the effect sizes of different treatments, and especially of drugs versus psychotherapy, are incomparable. Drug trials have excellent control conditions: double-blind placebo administration ensures that the only difference between the active and the control group is in the pharmacological action of the drug (apart from some possible bias due to unblinding).

Psychotherapy trials, on the other hand, often have very inadequate control conditions: many psychotherapy trials use wait-list controls and are also, by necessity, open or single-blind. Because of this, the effect size of psychotherapy versus wait-list is much larger than that of antidepressants versus placebo, at around 0.9 [543]. However, the effect size of psychotherapy versus other control conditions (e.g. care as usual, blinded pill placebo) in high-quality studies is much smaller, only around 0.22 [543, 79]. Direct comparisons between antidepressants and psychotherapy generally find that they are about as effective for depression and anxiety, with some exceptions (e.g. for dysthymia) [544].

Why are the effects of pharmacotherapy and of psychotherapy (compared to a good control condition) so small? In part, this is probably because some patients readily improve even in response to control conditions. In Chapter 9, for instance, we found that the pre-post effect size of placebo varied from 0.5 (for OCD) to 1.0 (for GAD). For depression, the pre-post effect size was 0.9 [69]. These pre-post effect sizes must be interpreted with caution, because it is impossible to separate true improvement from a mere regression to the mean effect [545], but they at least suggest that some patients do not require active treatment to improve.

In Chapters 9 and 10 I investigated baseline severity as a possible moderator of antidepressant efficacy, but I only found evidence that it was a significant moderator for GAD and panic disorder. This might imply that patients with mild GAD or panic disorder are more likely to improve spontaneously, although this must remain speculative: Chapter 10 (Figures 1 and 2) showed that the change in score in the placebo group actually increases with increasing severity (suggesting greater "spontaneous improvement"), but this may be due to regression to the mean or due to a floor effect at low levels of severity.

It remains unclear, therefore, which characteristics set apart those participants that are likely to respond adequately to control conditions, although we do have some knowledge of which patients show a relatively benign course even in the absence of treatment. Some of the characteristics that are associated with recovery include a short episode duration and mild symptoms [3, 546]. These characteristics motivate the practice of "watchful

waiting", which the Dutch guidelines recommend, for instance, for mild episodes of major depression with a duration of less than three months [74]. They are also applied equally to determine which patients can be assigned to low-intensity interventions like guided self-help, even though the (limited) available evidence suggests that these interventions are effective regardless of severity [547].

Among those who do not respond to control conditions or improve spontaneously, many also fail to respond to active treatment, whether antidepressants or psychotherapy, as reflected in the small effect sizes of these treatments. One explanation for this is that the etiology of depression and anxiety is likely to be heterogeneous *and* complex. These disorders can arise from purely organic causes (e.g. depression in Parkinson's disease [548]), but for most patients, they probably develop out of a complex interplay of genetic and other biological vulnerabilities, adverse early childhood experiences, inadequate or abusive parenting, trauma, poverty, personality disorders, childhood disorders like ADHD or autism, intellectual disability, and so on [549, 550].

Consequently, depression and anxiety are perhaps best conceptualized as the end stages of a very long illness process that more often than not has its roots in early childhood. It is therefore unsurprising that psychiatric treatments often prove inadequate in the face of this tangled knot of vulnerabilities. Worse, psychiatry is often forced to pit its treatments not just against these past vulnerabilities, but also against current, difficult to change, and deeply depressing, stressful, or anxiety-provoking life circumstances: unemployment, poverty, loneliness, abusive partners or family members, physical illness, caregiving responsibilities, single parenthood, and so on.

Seen in the light of our limited understanding of the brain [67], the deep roots of these disorders, and the adverse circumstances that many patients find themselves in, modest effect sizes are not surprising, especially considering that these treatment trials often only last six to twelve weeks. Our limited understanding of the brain also implies that revolutionary new treatments will probably be a long way off. Indeed, it could be argued that the treatment of depression and anxiety has not improved significantly since the development of the tricyclic antidepressants and benzodiazepines in the 1950s (although the SSRIs have better tolerability) and of cognitive (behavioral) therapy in the 1960s and 1970s. Instead, we have gained a diversity of approaches, all of which seem to be approximately equally effective [324, 544].

Although this proliferation of approaches has not resulted in the discovery of substantially more effective treatments, it could nevertheless be useful if some patients benefit from one particular treatment, while others benefit from a different treatment. The existence of such differential responses to treatments is a plausible hypothesis, given the heterogeneous etiology of depression and anxiety.

In the Sequenced Treatment Alternatives to Relieve Depression (STAR*D) trial, only 37% of participants attained remission in the first treatment step, consisting of up to 14

weeks of citalopram [551]. This is somewhat worse than our own finding that 51% of participants attained remission after twelve weeks of treatment (chapter 11), which may be because STAR*D had less strict inclusion and exclusion criteria (e.g. with regard to episode duration and comorbid disorders). For participants who stuck with treatment through multiple treatment failures, though, the cumulative remission rate was 67% in STAR*D.

While far from perfect, this nevertheless suggests that some patients can respond to second-line treatments even if they do not respond to an SSRI. The trick, then, is to achieve better matching of patients to treatments. Although we also need to develop effective new treatments for the 33% of participants who did not remit even after four treatment steps, this is a much more daunting undertaking than matching patients to the type and intensity of treatment that is most likely to work for them.

Better matching is also important to reduce treatment dropout. Depressed and anxious patients are likely to become discouraged after a treatment failure and may give up on treatment altogether. Dropout is strongly associated with poor outcomes: in STAR*D, for instance, 36% of non-remitters dropped out of treatment within 8 weeks, as compared to only 7% of eventual remitters [552]. Cause and effect probably go in both directions: participants who experience no benefit are likely to drop out; participants who drop out will experience no further benefit from treatment. It is therefore important to minimize the duration of ineffective treatment and to maximize the likelihood that the first attempted treatment is successful.

Precision psychiatry: making the best of what we have

One possible approach to maximizing the likelihood that the first attempted treatment is successful is, of course, to provide a high-intensity combination treatment to all patients. Addition of an atypical antipsychotic to an antidepressant leads to higher response rates in depression, for instance [472, 553], and combining psychotherapy and antidepressants is more effective than either treatment alone for both depression and anxiety [554]. However, the likelihood of success must be balanced with the likelihood of harm, which implies that patients should not receive more treatment than they require. Furthermore, patients should not continue receiving an ineffective treatment any longer than necessary.

In Part 2 of this thesis, I studied several clinical characteristics that might distinguish between patients who will benefit from treatment and those who will not. In particular, I examined the initial severity of symptoms as a predictor of response to antidepressants. In my IPD meta-analysis, I found evidence that antidepressant efficacy increased with increasing severity for GAD and panic disorder, but not for the other anxiety disorders. This implies that, for GAD and panic disorder, the benefits of antidepressants are rather small at low severity and as such, antidepressants may not be preferred as first-line

treatment.

The limited benefit of antidepressants compared to placebo should not be taken as encouragement to "do nothing", however, since taking part in the placebo group of a clinical trial is a fairly intensive intervention in and of itself. In support of the effectiveness of this intensive clinical management, it has been found that a greater frequency of trial visits is associated with a better clinical response in both the placebo and antidepressant group [555]. It remains to be seen whether the limited benefit of antidepressants in patients with mild GAD and panic disorder is because of spontaneous recovery (independent of trial participation) or because intensive clinical management is sufficient for these patients.

For patients with SAD, OCD, or PTSD, I did not find evidence that antidepressant efficacy depends upon initial severity. This implies that patients with mild to moderate disorders receive about as much benefit from antidepressants as patients with severe disorders. While this is encouraging news for patients who choose to take antidepressants, there may nevertheless be good reasons to reserve antidepressants (as a first-line treatment) for patients with more severe disorders. In patients with milder disorders, for instance, the burden of the disorder might not be proportional to the burden of antidepressant side effects.

It is also possible that patients with mild disorders experience more benefit from psychotherapy than patients with severe disorders, which could be an argument in favor of providing psychotherapy instead. The available evidence, however, suggests that the effectiveness of psychotherapy compared to antidepressants does not depend upon severity, although this evidence is derived from depression, not anxiety disorders [80]. Even so, it is possible that psychotherapy alone might suffice for patients with mild disorders, whereas patients with severe disorders might require both antidepressants and psychotherapy. Furthermore, for OCD, the evidence suggests that exposure and response prevention (ERP) and other CBT approaches are more effective than antidepressants and hence should generally be preferred regardless of severity [556].

In Chapter 11, I showed that early improvement, within the first two weeks, is predictive of later response or remission in patients with depression, but not so predictive that a blanket recommendation to change treatment for patients who show no early improvement should be implemented. Examining individual symptoms rather than the total score alone did not make a marked difference. At week 6, 26% of patients who showed no early improvement had responded, and by week 12, 38% had. Even patients who showed early worsening had a similar probability of response.

While this represents a considerable drop in the likelihood of a good outcome, these numbers are far from negligibly small. Furthermore, there is no evidence that switching antidepressants is associated with better outcomes than continuing the same antidepressant, either in the case of non-response after a conventional treatment period (six weeks or longer) [470] or in the case of non-improvement after two to four weeks [557, 558].

This suggests that lack of early improvement does not identify a subset of patients who are resistant to that particular antidepressant, but rather a subset of patients who are difficult to treat in general. Many of these patients may need more intensive treatment (e.g. augmentation or combination with psychotherapy), but since such intensified treatment is associated with a greater risk of harm, it seems premature to initiate it after just two weeks of monotherapy, in a patient group that still has about a 40% probability of responding to an antidepressant alone.

It might, however, be possible to identify a group of patients with an indication for early intensification of treatment by combining information about early improvement with baseline information about the previous course of the disorder, the presence of comorbid disorders or physical illnesses, circumstances that precipitated the disorder, and so on. In our IPD meta-analysis, the necessity of combining patient data from multiple trials precluded the incorporation of such information, because it was not consistently assessed across trials.

The main priority for future research, however, must be to identify characteristics that predict response to a specific treatment. Unfortunately, head-to-head trials of specific treatments are relatively scarce [544, 108], while there is an abundance of (markedly less useful) trials of antidepressants versus placebo [59], or of psychotherapy compared to a control treatment like wait-list or treatment-as-usual (TAU) [543]. The relative scarcity of head-to-head data may hamper the search for moderators of treatment efficacy, since fine-grained questions require relatively large amounts of data. An additional problem is that the study populations of many of the available trials (both head-to-head and otherwise) are highly selected and may not be particularly representative of the actual population of depressed and anxious people seeking treatments [475, 559, 560].

However, enough head-to-head trials have been done in the past few decades to at least begin to move from merely predicting outcome (regardless of treatment) to predicting response to specific treatments or at least treatment classes (e.g. SSRIs vs. SNRIs, CBT vs. antidepressants), especially for depression [81]. It is also becoming easier to access individual participant data from these trials. GSK was ahead of the curve in this regard by initiating Clinical Study Data Request (CSDR), which was used to access individual participant data for chapters 10 and 11 and which has now been joined by eleven other companies, including Lilly and Bayer. Other companies like Johnson & Johnson have agreed to a data sharing platform in collaboration with Yale, the so-called Yale Open Data Access (YODA) project [561], and Pfizer also makes clinical trial data available through its own portal [562].

Using these IPD is still hampered by lengthy application procedures, cumbersome access systems, and the inability to combine data from sponsors that participate in different platforms [563, 564], but since CSDR, the oldest of these platforms, was only initiated in 2013, it is not surprising that there is still room for improvement. Individual participant data from National Institute of Mental Health (NIMH)-sponsored trials, including

STAR*D, can also be requested [565], and IPD meta-analyses of psychotherapy trials have also been conducted [79, 80, 566], although these generally require the cooperation of the primary study authors.

In future research examining moderators of treatment efficacy, it will be important to adhere to best practices for predictive analytics – in particular, this means that the performance of predictive models must be tested in an independent data set in order to prevent over-fitting [567], a requirement that is not always met. Ideally, any hypotheses derived from such research should subsequently be tested in a prospective randomized trial, in which participants are assigned to their (model-predicted) "best" treatment or not, but this has the disadvantage that it would significantly delay implementation of these models in clinical practice.

Importantly, any model that predicts response to treatments that are similar in terms of harms, costs, and time investment (e.g. CBT versus another psychotherapy of similar intensity) only needs to be better than chance to be at least a little useful. However, we may require better performance of models that aim to predict which patients require intensive treatment (e.g. combination therapy), since mistakenly assigning a participant to the intensive treatment exposes them to potential harms and will result in greater costs.

Predictive models should, whenever possible, be based on information that is easy to collect in clinical practice or that is already routinely collected [68]. The work in chapters 9, 10 and 11 of this thesis suggests that information about symptoms alone is unlikely to yield very accurate predictions, but other clinical information about, for instance, illness history, family history, basic sociodemographic information other than age and gender, and the like, could easily be added. Should such models still prove inadequate, other information could be added that is more difficult to collect, for instance from blood samples or ecological momentary assessment. Even neuroimaging (for instance, functional magnetic resonance imaging (fMRI)) could be used, although given the expense of neuroimaging, it seems unlikely that this will soon be applied to the full population of treatment-naïve patients. Hence, predictive models that use difficult-to-collect information should probably be preferentially tested in a patient population with a relatively poor prognosis (e.g. those presenting in secondary care), for whom the application of such methods might actually be cost-effective.

Besides their immediate utility in clinical practice, these predictive models, if sufficiently refined, may also have other uses. First, patients who respond to the same treatment may also have the same or a similar underlying etiology. The heterogeneity of depression and anxiety is considered one important reason why progress in discovering the underlying causes of these disorders has been slow [568, 569]. Identifying more homogeneous patient subgroups with predictive models may therefore aid efforts to clarify the etiology of depression and anxiety, which in turn may help to refine predictive models, potentially setting in motion a virtuous cycle. Secondly, identifying patients who respond well to

various treatments will eventually also identify those patients who do not respond well to any treatment. Once we can identify patients who are truly treatment-resistant, efforts to develop new treatments can be focused on this patient group rather than on the entire group of depressed or anxious patients, many of whom are already adequately served by existing treatments.

Concluding remarks

This thesis aimed to bring the evidence regarding the treatment of depression and anxiety to light. To do so, I first examined the impact of reporting and citation biases on the evidence base. This effort has helped to elucidate how the apparent efficacy and safety of treatments, especially antidepressants, has been inflated. Having clarified the safety and efficacy of treatments, the next step is to determine who benefits from (which) treatment, and my thesis examined several clinical predictors that could be used for this. However, this area of research is still very much in its infancy. The increasing awareness of bias, coupled with the increasing availability of individual participant data, will hopefully allow for the continued development of evidence-based, precision psychiatry in future research.

Bibliography

[1] R C Kessler, P A Berglund, O Demler, R Jin, K R Merikangas, and E E Walters. Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication. *Arch Gen Psychiatry*, 62:593–603, 2005.

[2] American Psychiatric Association. *Diagnostic and statistical manual of mental disorders. 4th ed., text revision.* 2000.

[3] B W J H Penninx, W A Nolen, F Lamers, F G Zitman, J H Smit, and P Spinhoven et al. Two-year course of depressive and anxiety disorders: results from the Netherlands Study of Depression and Anxiety (NESDA). *J Affect Disord*, 133(1-2):76–85, 2011.

[4] R C Kessler, W T Chiu, O Demler, and E E Walters. Prevalence, severity, and comorbidity of 12-month DSM-IV disorders in the National Comorbidity Survey Replication. *Arch Gen Psychiatry*, 62:617–627, 2005.

[5] R F Krueger. The structure of common mental disorders. *Arch Gen Psychiatry*, 56(10):921–926, 1999.

[6] M M Fichter, N Quadflieg, U C Fischer, and G Kohlboeck. Twenty-five-year course and outcome in anxiety and depression in the Upper Bavarian Longitudinal Community Study. *Acta Psychiatr Scand*, 122(1):75–85, 2010.

[7] T E Moffitt, H Harrington, A Caspi, J Kim-Cohen, D Goldberg, A M Gregory, and R Poulton. Depression and generalized anxiety disorder: cumulative and sequential comorbidity in a birth cohort followed prospectively to age 32 years. *Arch Gen Psychiatry*, 64(6):651–60, 2007.

[8] E J Costello, S Mustillo, A Erkanli, G Keeler, and A Angold. Prevalence and development of psychiatric disorders in childhood and adolescence. *Arch Gen Psychiatry*, 60(8):837–44, 2003.

[9] J Ormel, D Raven, F van Oort, Catharina A Hartman, S A Reijneveld, R Veenstra, Wilma A M Vollebergh, J Buitelaar, F C Verhulst, and A J Oldehinkel. Mental health in Dutch adolescents: a TRAILS report on prevalence, severity, age of onset, continuity and co-morbidity of DSM disorders. *Psychol Med*, 45(02):345–360, 2015.

[10] R Gaspersz, F Lamers, J M Kent, A T F Beekman, J H Smit, A M van Hemert, R A Schoevers, and B W J H Penninx. Longitudinal predictive validity of the DSM-5 anxious distress specifier for clinical outcomes in a large cohort of patients with major depressive disorder. *J Clin Psychiatry*, 78(2):207–213, 2017.

[11] A Qaseem, M J Barry, and D Kansagara. Nonpharmacologic versus pharmacologic treatment of adult patients with major depressive disorder: a clinical practice guideline from the American College of Physicians. *Ann Intern Med*, 164:350–359, 2016.

[12] B Bandelow, L Sher, R Bunevicius, E Hollander, S Kasper, J Zohar, and H-J Möller. Guidelines for the pharmacological treatment of anxiety disorders, obsessive-compulsive disorder and posttraumatic stress disorder in primary care. *International Journal of Psychiatry in Clinical Practice*, 16(2):77–84, 2012.

[13] National Institute for Health and Clinical Excellence. Generalised anxiety disorder and panic disorder in adults: management (CG113). Technical report, 2011.

[14] National Institute for Health and Clinical Excellence. Obsessive-compulsive disorder and body dysmorphic disorder: treatment (CG31). Technical report, 2005.

[15] National Institute for Clinical Excellence. Post-traumatic stress disorder: management (CG26). Technical report, 2005.

[16] National Institute for Health and Care Excellence. Social anxiety disorder: recognition, assessment and treatment (CG159). Technical report, 2013.

[17] M E Franklin and E B Foa. Treatment of obsessive compulsive disorder. *Annu Rev Clin Psychol*, 7:229–243, 2011.

[18] D F Gros, N P Allan, and D D Szafranski. Movement towards transdiagnostic psychotherapeutic practices for the affective disorders. *Evid Based Ment Health*, 19(3):1–4, 2016.

[19] E H Turner, A M Matthews, E Linardatos, R A Tell, and R Rosenthal. Selective publication of antidepressant trials and its influence on apparent efficacy. *N Engl J Med*, 358(3):252–60, 2008.

[20] A M Roest, P de Jonge, C D Williams, Y A de Vries, R A Schoevers, and E H Turner. Reporting bias in clinical trials investigating the efficacy of second-generation antidepressants in the treatment of anxiety disorders: a report of 2 meta-analyses. *JAMA Psychiatry*, 72(5):500 – 510, 2015.

[21] M Huhn, M Tardy, L M Spineli, W Kissling, H Förstl, G Pitschel-Walz, C Leucht, M Samara, M Dold, J M Davis, and S Leucht. Efficacy of pharmacotherapy and psychotherapy for adult psychiatric disorders: a systematic overview of meta-analyses. *JAMA Psychiatry*, 71(6):706, 2014.

[22] F J Farach, L D Pruitt, J J Jun, Al B Jerud, L A Zoellner, and P P Roy-Byrne. Pharmacological treatment of anxiety disorders: Current treatments and future directions. *Journal of Anxiety Disorders*, 26(8):833–843, 2012.

[23] J Wong, A Motulsky, T Eguale, D L Buckeridge, M Abrahamowicz, and R Tamblyn. Treatment indications for antidepressants prescribed in primary care in Quebec, Canada, 2006 - 2015. *JAMA*, 315(20):2230–2231, 2016.

[24] T J Moore and D R Mattison. Adult utilization of psychiatric drugs and differences by sex, age, and race. *JAMA Intern Med*, 177(2):274 – 275, 2017.

[25] G Verweij and M Houben-van Herten. Bevolkingstrends 2013: Depressiviteit en antidepressiva in Nederland. Technical report, Centraal Bureau voor de Statistiek, 2013.

[26] L Cosgrove, S Krimsky, E E Wheeler, S M Peters, M Brodt, and A F Shaughnessy. Conflict of interest policies and industry relationships of guideline development group members: a cross-sectional study of clinical practice guidelines for depression. *Accountability in Research*, 24(2):99–115, 2016.

[27] H Melander, J Ahlqvist-Rastad, G Meijer, and B Beermann. Evidence b(i)ased medicine - selective reporting from studies sponsored by pharmaceutical industry: review of studies in new drug applications. *BMJ*, 326:1–5, 2003.

[28] J N Jureidini, L B McHenry, and P R Mansfield. Clinical trials and drug promotion: selective reporting of study 329. *Int J Risk Saf Med*, 20:73–81, 2008.

[29] J N Jureidini, J D Amsterdam, and L B McHenry. The citalopram CIT-MD-18 pediatric depression trial: Deconstruction of medical ghostwriting, data mischaracterisation and academic malfeasance. *Int J Risk Saf Med*, 28(1):33–43, 2016.

[30] R Keller, N D Ryan, M Strober, R G Klein, S P Kutcher, and B Birmaher et al. Efficacy of paroxetine in the treatment of adolescent major depression: a randomized, controlled trial. *J Am Acad Child Adolesc Psychiatry*, 40(7):762–772, 2001.

[31] M Parsons. Paroxetine in adolescent major depression. *J Am Acad Child Adolesc Psychiatry*, 41(4):364, 2002.

[32] J Jureidini and A Tonkin. Paroxetine in major depression. *J Am Acad Child Adolesc Psychiatry*, 42(5):514, 2003.

[33] J Le Noury, J M Nardo, D Healy, J Jureidini, M Raven, C Tufanaru, and E Abi-Jaoude. Restoring Study 329: efficacy and harms of paroxetine and imipramine in treatment of major depression in adolescence. *BMJ*, 351:h4320, 2015.

[34] D Fanelli. Negative results are disappearing from most disciplines and countries. *Scientometrics*, 90(3):891–904, 2012.

[35] E Driessen, S D Hollon, C L H Bockting, P Cuijpers, and E H Turner. Does publication bias inflate the apparent efficacy of psychological treatment for major depressive disorder? A systematic review and meta-analysis of US National Institutes of Health-funded trials. *PLOS ONE*, 10(9):e0137864, 2015.

[36] T D Sterling. Publication decisions and their possible effects on inferences drawn from

tests of significance–or vice versa. *Journal of the American Statistical Association*, 54(285):30, 1959.

[37] S Hopewell, K Loudon, M J Clarke, A D Oxman, and K Dickersin. Publication bias in clinical trials due to statistical significance or direction of trial results. *Cochrane Database Syst Rev*, (1), 2009.

[38] K Dwan, C Gamble, P R Williamson, and J J Kirkham. Systematic review of the empirical evidence of study publication bias and outcome reporting bias - an updated review. *PLOS ONE*, 8(7):e66844, 2013.

[39] S Mathieu, I Boutron, D Moher, D G Altman, and P Ravaud. Comparison of registered and published primary outcomes in randomized controlled trials. *JAMA*, 302(9):977, 2009.

[40] I Boutron, S Dutton, P Ravaud, and D G Altman. Reporting and interpretation of randomized controlled trials with statistically nonsignificant results for primary outcomes. *JAMA*, 303(20):2058–64, 2010.

[41] I Boutron, D G Altman, S Hopewell, F Vera-Badillo, I Tannock, and P Ravaud. Impact of spin in the abstracts of articles reporting results of randomized controlled trials in the field of cancer: the SPIIN randomized controlled trial. *Journal of Clinical Oncology*, 32(36):4120–4126, 2014.

[42] J-F Etter and J Stapleton. Citations to trials of nicotine replacement therapy were biased toward positive results and high-impact-factor journals. *J Clin Epidemiol*, 62:831–837, 2009.

[43] M Callaham, R L Wears, and E Weber. Journal prestige, publication bias, and other characteristics associated with citation of published studies in peer-reviewed journals. *JAMA*, 287(21):2847–2850, 2002.

[44] S A Greenberg. How citation distortions create unfounded authority: analysis of a citation network. *BMJ*, 339:b2680, 2009.

[45] P Nieminen, G Rucker, J Miettunen, J Carpenter, and M Schumacher. Statistically significant papers in psychiatry were cited more often than others. *J Clin Epidemiol*, 60:939–946, 2007.

[46] F Song, S Parekh, L Hooper, Y K Loke, J Ryder, A J Sutton, C Hing, C S Kwok, C Pang, and I Harvey. Dissemination and publication of research findings: An updated review of related biases. *Health Technology Assessment*, 14(8):1–220, 2010.

[47] U Jonsson, I Alaie, T Parling, and F K Arnberg. Reporting of harms in randomized controlled trials of psychological interventions for mental and behavioral disorders: a review of current practice. *Contemp Clin Trials*, 38(1):1–8, 2014.

[48] B Vaughan, M H Goldstein, M Alikakos, L J Cohen, and M J Serby. Frequency of reporting of adverse events in randomized controlled trials of psychotherapy vs. psychopharmacotherapy. *Compr Psychiatry*, 55(4):849–55, 2014.

[49] M Linden and M-L Schermuly-Haupt. Definition, assessment and rate of psychotherapy side effects. *World Psychiatry*, 13(3):306–309, 2014.

[50] D J Nutt and M Sharpe. Uncritical positive regard? Issues in the efficacy and safety of psychotherapy. *J Psychopharmacol*, 22(1):3–6, 2008.

[51] M J Crawford, L Thana, L Farquharson, L Palmer, E Hancock, P Bassett, J Clarke, and G D Parry. Patient experience of negative effects of psychological treatment: results of a national survey. *Br J Psychiatry*, 208(3):260–265, 2016.

[52] A Rozental, A Kottorp, J Boettcher, G Andersson, and P Carlbring. Negative effects of psychological treatments: An exploratory factor analysis of the Negative Effects Questionnaire for monitoring and reporting adverse and unwanted events. *PLOS ONE*, 11(6):e0157503, 2016.

[53] P N Papanikolaou, R Churchill, K Wahlbeck, and J P A Ioannidis. Safety reporting in randomized trials of mental health interventions. *Am J Psychiatry*, 161(9):1692–7, 2004.

[54] John P A Ioannidis and Joseph Lau. Completeness of safety reporting in randomized trials. *JAMA*, 285(4):437, jan 2001.

[55] B Wieseler, N Wolfram, N McGauran, M F Kerekes, V Vervölgyi, P Kohlepp, M Kamphuis, and U Grouven. Completeness of reporting of patient-relevant clinical trial outcomes: comparison of unpublished clinical study reports with publicly available data. *PLOS Med*, 10(10):e1001526, 2013.

[56] Y K Loke and S Derry. Reporting of adverse drug reactions in randomised controlled trials - a systematic survey. *BMC Clinical Pharmacology*, 1:3, 2001.
[57] E Maund, B Tendal, A Hróbjartsson, K J Jorgensen, A Lundh, J Schroll, and P C Gøtzsche. Benefits and harms in clinical trials of duloxetine for treatment of major depressive disorder: comparison of clinical study reports, trial registries, and publications. *BMJ*, 348:g3510, 2014.
[58] S Hughes, D Cohen, and R Jaggi. Differences in reporting serious adverse events in industry sponsored clinical trial registries and journal articles on antidepressant and antipsychotic drugs: a cross-sectional study. *BMJ Open*, 4(7):e005535, 2014.
[59] J Undurraga and R J Baldessarini. Randomized, placebo-controlled trials of antidepressants for acute major depression: thirty-year meta-analytic review. *Neuropsychopharmacology*, 37(4):851–864, 2012.
[60] M J Taylor, S Sen, and Z Bhagwagar. Antidepressant response and the serotonin transporter gene-linked polymorphic region. *Biol Psychiatry*, 68(6):536–543, 2010.
[61] S Porcelli, C Fabbri, and A Serretti. Meta-analysis of serotonin transporter gene promoter polymorphism (5-HTTLPR) association with antidepressant efficacy. *Eur Neuropsychopharmacol*, 22(4):239–58, 2012.
[62] J Flint and M R Munafò. Candidate and non-candidate genes in behavior genetics. *Curr Opin Neurobiol*, 23(1):57–61, 2013.
[63] P F Sullivan. Spurious genetic associations. *Biol Psychiatry*, 61(10):1121–6, 2007.
[64] J P Simmons, L D Nelson, and U Simonsohn. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol Sci*, 22(11):1359–66, 2011.
[65] J P A Ioannidis, R Tarone, and J K McLaughlin. The false-positive to false-negative ratio in epidemiologic studies. *Epidemiology*, 22(4):450–6, 2011.
[66] R Uher, K E Tansey, M Rietschel, N Henigsberg, W Maier, and O Mors et al. Common genetic variation and antidepressant efficacy in major depressive disorder: A meta-analysis of three genome-wide pharmacogenetic studies. *Am J Psychiatry*, 170(2):207–217, 2013.
[67] S Kapur, A G Phillips, and T R Insel. Why has it taken so long for biological psychiatry to develop clinical tests and what to do about it? *Mol Psychiatry*, 17:1174–1179, 2012.
[68] T Hahn, Andrew A Nierenberg, and S Whitfield-Gabrieli. Predictive analytics in mental health: applications, guidelines, challenges and perspectives. *Mol Psychiatry*, 22(1):37–43, 2016.
[69] I Kirsch, B J Deacon, T B Huedo-Medina, A Scoboria, T J Moore, and B T Johnson. Initial severity and antidepressant benefits: a meta-analysis of data submitted to the Food and Drug Administration. *PLOS Med*, 5(2):e45, 2008.
[70] A Khan, R M Leventhal, S R F Khan, and W A Brown. Severity of depression and response to antidepressants and placebo: an analysis of the Food and Drug Administration database. *J Clin Psychopharmacol*, 22(1):40–45, 2002.
[71] A Khan, A E Brodhead, R L Kolts, and W A Brown. Severity of depressive symptoms and response to antidepressants and placebo in antidepressant trials. *J Psychiatr Res*, 39(2):145–50, 2005.
[72] J C Fournier, R J Derubeis, S D Hollon, S Dimidjian, J D Amsterdam, R C Shelton, and J Fawcett. Antidepressant drug effects and depression severity. *JAMA*, 303(1):47–53, 2010.
[73] National Collaborating Centre for Mental Health. Depression: the NICE guideline on the treatment and management of depression in adults. Technical report, National Center for Clinical Excellence, 2010.
[74] J Spijker, Claudi L H Bockting, J A C Meeuwissen, I M van Vliet, P M G Emmelkamp, M L M Hermens, and Anton J L M van Balkom. Multidisciplinaire richtlijn depressie (3e revisie, 2013). Technical report, Trimbos-instituut, Utrecht, 2013.
[75] J Rabinowitz, N Werbeloff, F S Mandel, L Marangell, and S Kapur. Initial depression severity and response to antidepressants v . placebo: patient-level data analysis from 34 randomised controlled trials. *Br J Psychiatry*, 209(5):427–428, 2016.
[76] R D Gibbons, K Hur, C H Brown, J M Davis, and J J Mann. Benefits from antidepressants: synthesis of 6-week patient-level outcomes from double-blind placebo-controlled randomized trials of fluoxetine and venlafaxine. *Arch Gen Psychiatry*, 69(6):572–9, 2012.

[77] M A Sugarman, A M Loree, B B Baltes, E R Grekin, and I Kirsch. The efficacy of paroxetine and placebo in treating anxiety and depression: a meta-analysis of change on the Hamilton rating scales. *PLOS ONE*, 9(8):e106337, 2014.

[78] D L Ackerman and S Greenland. Multivariate meta-analysis of controlled drug studies for obsessive-compulsive disorder. *J Clin Psychopharmacol*, 22(3):309–17, 2002.

[79] T A Furukawa, E S Weitz, S Tanaka, S D Hollon, S G Hofmann, and G Andersson et al. Initial severity of depression and efficacy of cognitive–behavioural therapy: individual-participant data meta-analysis of pill-placebo-controlled trials. *Br J Psychiatry*, 210(3):190–196, 2017.

[80] E S Weitz, S D Hollon, J Twisk, A van Straten, M J H Huibers, and D David et al. Baseline depression severity as moderator of depression outcomes between cognitive behavioral therapy vs pharmacotherapy. *JAMA Psychiatry*, 72(11):1102 – 1109, 2015.

[81] G E Simon and R H Perlis. Personalized medicine for depression: Can we match patients with treatments? *Am J Psychiatry*, 167(12):1445–1455, 2010.

[82] R H Perlis. A clinical risk stratification tool for predicting treatment resistance in major depressive disorder. *Biol Psychiatry*, 74(1):7–14, 2013.

[83] A M Chekroud, R J Zotti, Z Shehzad, R Gueorguieva, M K Johnson, M H Trivedi, T D Cannon, J H Krystal, and P R Corlett. Cross-trial prediction of treatment outcome in depression: a machine learning approach. *Lancet Psychiatry*, 3(3):243–250, 2016.

[84] R J DeRubeis, Z D Cohen, N R Forand, J C Fournier, L A Gelfand, and L Lorenzo-Luaces. The Personalized Advantage Index: translating research on prediction into individualized treatment recommendations. A demonstration. *PLOS ONE*, 9(1):e83875, 2014.

[85] M J H Huibers, Z D Cohen, L H J M Lemmens, A Arntz, F P M L Peeters, P Cuijpers, and R J DeRubeis. Predicting optimal outcomes in cognitive therapy or interpersonal psychotherapy for depressed individuals using the Personalized Advantage Index approach. *PLoS ONE*, 10(11):1–16, 2015.

[86] American Psychiatric Association. Practice guideline for the treatment of patients with major depressive disorder. Technical report, 2010.

[87] A Szegedi, W T Jansen, A P P Van Willigenburg, E Van Der Meulen, H H Stassen, and M E Thase. Early improvement in the first 2 weeks as a predictor of treatment outcome in patients with major depressive disorder: A meta-analysis including 6562 patients. *J Clin Psychiatry*, 70(3):344–353, 2009.

[88] E Jakubovski and M H Bloch. Prognostic subgroups for citalopram response in the STAR*D trial. *J Clin Psychiatry*, 75(7):738–747, 2014.

[89] M Rynn, S Khalid-Khan, J F Garcia-Espana, B Etemad, and K Rickels. Early response and 8-week treatment outcome in GAD. *Depress Anxiety*, 23(8):461–465, 2006.

[90] D S Baldwin, E Schweizer, Y Xu, and G Lyndon. Does early improvement predict endpoint response in patients with generalized anxiety disorder (GAD) treated with pregabalin or venlafaxine XR? *Eur Neuropsychopharmacol*, 22(2):137–142, 2012.

[91] M Albus, Y Lecrubier, W Maier, R Buller, R Rosenberg, and H Hippius. Drug treatment of panic disorder: early response to treatment as a predictor of final outcome. *Acta Psychiatr Scand*, 82(5):359–365, 1990.

[92] M H Pollack, M H Rapaport, R Fayyad, M W Otto, A A Nierenberg, and C M Clary. Early improvement predicts endpoint remission status in sertraline and placebo treatments of panic disorder. *J Psychiatr Res*, 36(4):229–236, 2002.

[93] Mark H Pollack, Susan G Kornstein, Melissa E Spann, Paul Crits-Christoph, Joel Raskin, and James M Russell. Early improvement during duloxetine treatment of generalized anxiety disorder predicts response and remission at endpoint. *J Psychiatr Res*, 42(14):1176–1184, 2008.

[94] D S Baldwin, D J Stein, O T Dolberg, and B Bandelow. How long should a trial of escitalopram treatment be in patients with major depressive disorder, generalised anxiety disorder or social anxiety disorder? An exploration of the randomised controlled trial database. *Hum Psychopharmacol*, 24(4):269–275, 2009.

[95] R Uher, O Mors, M Rietschel, A Rajewska-Rager, A Petrovic, A Zobel, N Henigsberg, J Mendlewicz, K J Aitchison, A Farmer, and P McGuffin. Early and delayed onset of response to antidepressants in individual trajectories of change during treatment of major

depression: A secondary analysis of data from the genome-based therapeutic drugs for depression (GENDEP) study. *J Clin Psychiatry*, 72(11):1478–1484, 2011.

[96] E I Fried and R M Nesse. Depression sum-scores don't add up: why analyzing specific depression symptoms is essential. *BMC Med*, 13(1):72, 2015.

[97] E I Fried and R M Nesse. Depression is not a consistent syndrome: An investigation of unique symptom patterns in the STAR*D study. *J Affect Disord*, 172C:96–102, 2014.

[98] H Sakurai, H Uchida, T Abe, S Nakajima, T Suzuki, B G Pollock, Y Sato, and M Mimura. Trajectories of individual symptoms in remitters versus non-remitters with depression. *J Affect Disord*, 151(2):506–513, 2013.

[99] H Tokuoka, H Takahashi, A Ozeki, A Kuga, A Yoshikawa, T Tsuji, and M M Wohlreich. Trajectories of depression symptom improvement and associated predictor analysis: An analysis of duloxetine in double-blind placebo-controlled trials. *J Affect Disord*, 196:171–180, 2016.

[100] K Funaki, S Nakajima, T Suzuki, M Mimura, and H Uchida. Early improvements in individual symptoms to predict later remission in major depressive disorder treated with mirtazapine. *J Clin Pharmacol*, 56(9):1111–1119, jan 2016.

[101] D Eyding and M Lelgemann. Reboxetine for acute treatment of major depression: systematic review and meta-analysis of published and unpublished placebo and selective serotonin reuptake inhibitor controlled trials. *BMJ*, 341:c4737, 2010.

[102] J P T Higgins and S Green, editors. *Cochrane Handbook for Systematic Reviews of Interventions*. The Cochrane Collaboration, 5 edition, 2011.

[103] A-W Chan, A Hróbjartsson, M T Haahr, P C Gøtzsche, and D G Altman. Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. *JAMA*, 291(20):2457–2465, 2004.

[104] E H Turner, D Knoepflmacher, and L Shapley. Publication bias in antipsychotic trials: an analysis of efficacy comparing the published literature to the US Food and Drug Administration database. *PLOS Med*, 9(3):e1001189, 2012.

[105] Erick H Turner. A taxpayer-funded clinical trials registry and results database. *PLOS Med*, 1(3):e60, dec 2004.

[106] E H Turner. How to access and process FDA drug approval packages for use in research. *BMJ*, (347):f5992, 2013.

[107] John P A Ioannidis. Effectiveness of antidepressants: an evidence myth constructed from a thousand randomized trials? *Philosophy, Ethics, and Humanities in Medicine*, 3:14, 2008.

[108] A Cipriani, T A Furukawa, G Salanti, J R Geddes, J P T Higgins, R Churchill, N Watanabe, A Nakagawa, I M Omori, H McGuire, M Tansella, and C Barbui. Comparative efficacy and acceptability of 12 new-generation antidepressants: a multiple-treatments meta-analysis. *Lancet*, 373(9665):746–58, 2009.

[109] M Olfson and S C Marcus. National patterns in antidepressant medication treatment. *Arch Gen Psychiatry*, 66(8):848–856, 2009.

[110] A J Baxter, K M Scott, T Vos, and H A Whiteford. Global prevalence of anxiety disorders: a systematic review and meta-regression. *Psychol Med*, 43(5):897–910, 2013.

[111] J R T Davidson. First-line pharmacotherapy approaches for generalized anxiety disorder. *J Clin Psychiatry*, 70(suppl 2):25–31, 2009.

[112] N M Batelaan, A J L M van Balkom, and D J Stein. Evidence-based pharmacotherapy of panic disorder: an update. *Int J Neuropsychopharmacol*, 15(3):403–415, 2011.

[113] C Blanco, L B Bragdon, F R Schneier, and M R Liebowitz. The evidence-based pharmacotherapy of social anxiety disorder. *Int J Neuropsychopharmacol*, 16(1):235–49, 2013.

[114] J C Ipser and D J Stein. Evidence-based pharmacotherapy of post-traumatic stress disorder (PTSD). *Int J Neuropsychopharmacol*, 15(06):825–840, 2012.

[115] N A Fineberg, A Brown, S Reghunandanan, and I Pampaloni. Evidence-based pharmacotherapy of obsessive-compulsive disorder. *Int J Neuropsychopharmacol*, 15(08):1173–1191, 2012.

[116] R B Hidalgo, L A Tupler, and J R T Davidson. An effect-size analysis of pharmacologic treatments for generalized anxiety disorder. *J Psychopharmacol*, 21(8):864–72, 2007.

[117] F Kapczinski, J J S S dos Santos Souza, A A B C Batista Miralha da Cunha, and R R S

Schmitt. Antidepressants for generalised anxiety disorder (GAD). *Cochrane Database Syst Rev*, 2:CD003592, 2003.

[118] M W Otto, K S Tuby, R A Gould, R Y McLean, and Mark H Pollack. An effect-size analysis of the relative efficacy and tolerability of serotonin selective reuptake inhibitors for panic disorder. *Am J Psychiatry*, 158(12):1989–92, 2001.

[119] C Andrisano, A Chiesa, and A Serretti. Newer antidepressants and panic disorder: a meta-analysis. *Int Clin Psychopharmacol*, 28(1):33–45, 2013.

[120] D W Hedges, B L Brown, D A Shwalb, K Godfrey, and A M Larcher. The efficacy of selective serotonin reuptake inhibitors in adult social anxiety disorder: a meta-analysis of double-blind, placebo-controlled trials. *J Psychopharmacol*, 21(1):102–11, 2007.

[121] G B De Menezes, E S F Coutinho, L F Fontenelle, P Vigne, I Figueira, and M Versiani. Second-generation antidepressants in social anxiety disorder: Meta-analysis of controlled clinical trials. *Psychopharmacology*, 215(1):1–11, 2011.

[122] D J Stein, J C Ipser, and S Seedat. Pharmacotherapy for post traumatic stress disorder (PTSD). *Cochrane Database Syst Rev*, 1:CD002795, 2006.

[123] G M Soomro, D G Altman, S Rajagopal, and M Oakley Browne. Selective serotonin re-uptake inhibitors (SSRIs) versus placebo for obsessive compulsive disorder (OCD). *Cochrane Database Syst Rev*, 1:CD001765, 2008.

[124] M W Lipsey and D Wilson. *Practical Meta-analysis (Applied Social Research Methods)*. Sage Publications, 2001.

[125] J P T Higgins and S G Thompson. Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, 21(11):1539–1558, 2002.

[126] K Rickels, M Rynn, M Iyengar, and D Duff. Remission of generalized anxiety disorder: a review of the paroxetine clinical trials database. *J Clin Psychiatry*, 67:41–47, 2006.

[127] D J Stein, J Davidson, S Seedat, and K Beebe. Paroxetine in the treatment of post-traumatic stress disorder: pooled analysis of placebo-controlled studies. *Expert Opin Pharmacother*, 4(10):1829–1838, 2003.

[128] J I Sheikh, P Londborg, C M Clary, and R Fayyad. The efficacy of sertraline in panic disorder: combined results from two fixed-dose studies. *Int Clin Psychopharmacol*, 15(6):335–342, 2000.

[129] P D Londborg, R Wolkow, W T Smith, E DuBoff, D England, J Ferguson, M Rosenthal, and C Weise. Sertraline in the treatment of panic disorder. A multi-site, double- blind, placebo-controlled, fixed-dose investigation. *Br J Psychiatry*, 173(1):54–60, 1998.

[130] W K Goodman, A Bose, and Q Wang. Treatment of generalized anxiety disorder with escitalopram: Pooled results from double-blind, placebo-controlled trials. *J Affect Disord*, 87(2-3):161–167, 2005.

[131] J R T Davidson, A Bose, A Korotzer, and H Zheng. Escitalopram in the treatment of generalized anxiety disorder: Double-blind, placebo controlled, flexible-dose study. *Depress Anxiety*, 19(4):234–240, 2004.

[132] K Rickels, R Zaninelli, J McCaffery, K Bellew, M Iyengar, and D Sheehan. Paroxetine treatment of generalized anxiety disorder: a double-blind, placebo-controlled study. *Am J Psychiatry*, 160:749 – 756, 2003.

[133] M H Pollack, R Zaninelli, A Goddard, J P McCafferty, K M Bellew, D B Burnham, and M K Iyengar. Paroxetine in the treatment of generalized anxiety disorder: results of a placebo-controlled, flexible-dosage trial. *J Clin Psychiatry*, 62(5):350–7, 2001.

[134] H Koponen, C Allgulander, J Erickson, E Dunayevich, Y Pritchett, M J Detke, S G Ball, and J M Russell. Efficacy of duloxetine for the treatment of generalized anxiety disorder: implications for primary care physicians. *Prim Care Companion J Clin Psychiatry*, 9(2):100–7, 2007.

[135] M Rynn, J Russell, J Erickson, M J Detke, S Ball, J Dinkel, K Rickels, and J Raskin. Efficacy and safety of duloxetine in the treatment of generalized anxiety disorder: a flexible-dose, progressive-titration, placebo-controlled trial. *Depress Anxiety*, 25(3):182–189, 2008.

[136] J Hartford, S Kornstein, M Liebowitz, T Pigott, J Russell, M Detke, D Walker, S Ball, E Dunayevich, J Dinkel, and J Erickson. Duloxetine as an SNRI treatment for generalized anxiety disorder: results from a placebo and active-controlled trial. *Int Clin Psychopharmacol*, 22(3):167–74, 2007.

[137] K Rickels, M H Pollack, D V Sheehan, and J T Haskins. Efficacy of extended-release venlafaxine in nondepressed outpatients with generalized anxiety disorder. *Am J Psychiatry*, 157(6):968–974, 2000.

[138] J R T Davidson, R L DuPont, D Hedges, and J T Haskins. Efficacy, safety, and tolerability of venlafaxine extended release and buspirone in outpatients with generalized anxiety disorder. *J Clin Psychiatry*, 60(8):528–535, 1999.

[139] J C Ballenger, D E Wheadon, M Steiner, W Bushnell, and I P Gergel. Double-blind, fixed-dose, placebo-controlled study of paroxetine in the treatment of panic disorder. *Am J Psychiatry*, 155(1):36–42, 1998.

[140] S Oehrberg, P E Christiansen, K Behnke, A L Borup, B Severin, J Soegaard, H Calberg, R Judge, J K Ohrstrom, and P M Manniche. Paroxetine in the treatment of panic disorder. A randomised, double-blind, placebo-controlled study. *Br J Psychiatry*, 167:374–379, 1995.

[141] Y Lecrubier, A Bakker, G Dunbar, and R Judge. A comparison of paroxetine, clomipramine and placebo in the treatment of panic disorder. *Acta Psychiatr Scand*, 95(2):145–52, 1997.

[142] D V Sheehan, D B Burnham, M K Iyengar, and P Perera. Efficacy and tolerability of controlled-release paroxetine in the treatment of panic disorder. *J Clin Psychiatry*, 66(1):34–40, 2005.

[143] R B Pohl, R M Wolkow, and C M Clary. Sertraline in the treatment of panic disorder: a double-blind multicenter trial. *Am J Psychiatry*, 155(9):1189–95, 1998.

[144] M H Pollack, M W Otto, J J Worthington, G G Manfro, and R Wolkow. Sertraline in the treatment of panic disorder: a flexible-dose multicenter trial. *Arch Gen Psychiatry*, 55(11):1010–1016, 1998.

[145] M H Pollack, U Lepola, H Koponen, N M Simon, J J Worthington, G Emilien, E Tzanis, E Salinas, T Whitaker, and B Gao. A double-blind study of the efficacy of venlafaxine extended-release, paroxetine, and placebo in the treatment of panic disorder. *Depress Anxiety*, 24(1):1–14, 2007.

[146] M Pollack, R Mangano, R Entsuah, E Tzanis, N M Simon, and Y Zhang. A randomized controlled trial of venlafaxine ER and paroxetine in the treatment of outpatients with panic disorder. *Psychopharmacology*, 194(2):233–42, 2007.

[147] M R Liebowitz, G Asnis, R Mangano, and E Tzanis. A double-blind, placebo-controlled, parallel-group, flexible-dose study of venlafaxine extended release capsules in adult outpatients with panic disorder. *J Clin Psychiatry*, 70(4):550–561, 2009.

[148] J Bradwejn, A Ahokas, D J Stein, E Salinas, G Emilien, and T Whitaker. Venlafaxine extended-release capsules in panic disorder: Flexible-dose, double-blind, placebo-controlled study. *Br J Psychiatry*, 187(10):352–359, 2005.

[149] J Davidson, J Yaryura-Tobias, R DuPont, L Stallings, L M Barbato, R G van der Hoop, and D Li. Fluvoxamine-controlled release formulation for the treatment of generalized social anxiety disorder. *J Clin Psychopharmacol*, 24(2):118–125, 2004.

[150] H G M Westenberg, D J Stein, H Yang, D Li, and L M Barbato. A double-blind placebo-controlled study of controlled release fluvoxamine for the treatment of generalized social anxiety disorder. *J Clin Psychopharmacol*, 24(1):49–55, 2004.

[151] D Baldwin, J Bobes, D J Stein, I Scharwachter, and M Faure. Paroxetine in social phobia/social anxiety disorder: randomised, double- blind, placebo-controlled study. *Br J Psychiatry*, 175(2):120–126, 1999.

[152] M B Stein, M R Liebowitz, R B Lydiard, C D Pitts, W Bushnell, and I Gergel. Paroxetine treatment of generalized social phobia (social anxiety disorder): a randomized controlled trial. *JAMA*, 280(8):708–713, 1998.

[153] M R Liebowitz, M B Stein, M Tancer, D Carpenter, R Oakes, and C D Pitts. A randomized, double-blind, fixed-dose comparison of paroxetine and placebo in the treatment of generalized social anxiety disorder. *J Clin Psychiatry*, 63:66 – 74, 2002.

[154] U Lepola, B Bergtholdt, J St Lambert, K L Davy, and L Ruggiero. Controlled-release paroxetine in the treatment of patients with social anxiety disorder. *J Clin Psychiatry*, 65(2):222–9, 2004.

[155] M R Liebowitz, N A DeMartinis, K Weihs, P D Londborg, W T Smith, H Chung, R Fayyad, and C M Clary. Efficacy of sertraline in severe generalized social anxiety disor-

der: results of a double-blind, placebo-controlled study. *J Clin Psychiatry*, 64(7):785–92, 2003.

[156] M A Van Ameringen, R M Lane, J R Walker, R C Bowen, P R Chokka, E M Goldner, D G Johnston, Y J Lavallee, S Nandy, J C Pecknold, V Hadrava, and R P Swinson. Sertraline treatment of generalized social phobia: a 20-week, double-blind, placebo-controlled study. *Am J Psychiatry*, 158(2):275–281, 2001.

[157] S Blomhoff, T T Haug, K Hellström, I Holme, M Humble, H P Madsbu, and J E Wold. Randomised controlled general practice trial of sertraline, exposure therapy and combined treatment in generalised social phobia. *Br J Psychiatry*, 179(1):23–30, 2001.

[158] M R Liebowitz, R M Mangano, J Bradwejn, and G Asnis. A randomized controlled trial of venlafaxine extended release in generalized social anxiety disorder. *J Clin Psychiatry*, 66(2):238–247, 2005.

[159] K Rickels, R Mangano, and A Khan. A double-blind, placebo-controlled study of a flexible dose of venlafaxine ER in adult outpatients with generalized social anxiety disorder. *J Clin Psychopharmacol*, 24(5):488–496, 2004.

[160] M J Friedman, C R Marmar, D G Baker, C S Sikes, and G M Farfel. Randomized, double-blind comparison of sertraline and placebo for posttraumatic stress disorder in a Department of Veterans Affairs setting. *J Clin Psychiatry*, 68(5):711–720, 2008.

[161] J R Davidson, B O Rothbaum, B A van der Kolk, C R Sikes, and G M Farfel. Multicenter, double-blind comparison of sertraline and placebo in the treatment of posttraumatic stress disorder. *Arch Gen Psychiatry*, 58(5):485–92, 2001.

[162] K Brady, T Pearlstein, G M Asnis, D Baker, B Rothbaum, C R Sikes, and G M Farfel. Efficacy and safety of sertraline treatment of posttraumatic stress disorder. *JAMA*, 283(14):1837, 2000.

[163] R D Marshall, K L Beebe, M Oldham, and R Zaninelli. Efficacy and safety of paroxetine treatment for chronic PTSD: A fixed-dose, placebo-controlled study. *Am J Psychiatry*, 158(12):1982–1988, 2001.

[164] P Tucker, R Zaninelli, R Yehuda, L Ruggiero, K Dillingham, and C D Pitts. Paroxetine in the treatment of chronic posttraumatic stress disorder: results of a placebo-controlled, flexible-dosage trial. *J Clin Psychiatry*, 62(11):860–868, 2001.

[165] G D Tollefson, A H Rampey, J H Potvin, M A Jenike, A J Rush, R A Dominguez, L M Koran, M K Shear, W Goodman, and L A Genduso. A multicenter investigation of fixed-dose fluoxetine in the treatment of obsessive-compulsive disorder. *Arch Gen Psychiatry*, 51(7):559–567, 1994.

[166] S A Montgomery, A McIntyre, M Osterheider, P Sarteschi, W Zitterl, J Zohar, M Birkett, and A J Wood. A double-blind, placebo-controlled study of fluoxetine in patients with DSM-III-R obsessive-compulsive disorder. The Lilly European OCD Study Group. *Eur Neuropsychopharmacol*, 3(2):143–52, 1993.

[167] W K Goodman, M J Kozak, M Liebowitz, and K L White. Treatment of obsessive-compulsive disorder with fluvoxamine: a multicentre, double-blind, placebo-controlled trial. *Int Clin Psychopharmacol*, 11(1):21–9, 1996.

[168] E Hollander, L M Koran, W K Goodman, J H Greist, P T Ninan, H Yang, D Li, and L M Barbato. A double-blind, placebo-controlled study of the efficacy and safety of controlled-release fluvoxamine in patients with obsessive-compulsive disorder. *J Clin Psychiatry*, 64(6):640–7, 2003.

[169] E Hollander, A Allen, M Steiner, D E Wheadon, R Oakes, and D B Burnham. Acute and long-term treatment and prevention of relapse of obsessive-compulsive disorder with paroxetine. *J Clin Psychiatry*, 64(9):1113–1121, 2003.

[170] J Zohar and R Judge. Paroxetine versus clomipramine in the treatment of obsessive-compulsive disorder. *Br J Psychiatry*, 169(4):468–474, 1996.

[171] G Chouinard, W Goodman, J Greist, M Jenike, S Rasmussen, K White, E Hackett, M Gaffney, and P A Bick. Results of a double-blind placebo controlled trial of a new serotonin uptake inhibitor, sertraline, in the treatment of obsessive-compulsive disorder. *Psychopharmacol Bull*, 26(3):279–284, 1990.

[172] J Greist, G Chouinard, E DuBoff, A Halaris, S W Kim, L Koran, M Liebowitz, R B Lydiard, S Rasmussen, and K White. Double-blind parallel comparison of three dosages

of sertraline and placebo in outpatients with obsessive-compulsive disorder. *Arch Gen Psychiatry*, 52(4):289–95, 1995.

[173] M H Kronig, J Apter, G Asnis, A Bystritsky, G Curtis, J Ferguson, R Landbloom, D Munjack, R Riesenberg, D Robinson, P Roy-Byrne, K Phillips, and I J Du Pont. Placebo-controlled, multicenter study of sertraline treatment for obsessive-compulsive disorder. *J Clin Psychopharmacol*, 19(2):172–6, 1999.

[174] Gl I Spielmans, T L Biehn, and D L Sawrey. A case study of salami slicing: pooled analyses of duloxetine for depression. *Psychother Psychosom*, 79(2):97–106, 2010.

[175] K J Thaler, L C Morgan, M Van Noord, D E Jonas, M S McDonagh, K Peterson, A Glechner, and G Gartlehner. A case study of pooled-studies publications indicated potential for both valuable information and bias. *J Clin Epidemiol*, 66(10):1082–1092, 2013.

[176] Erick H Turner. Publication bias, with a focus on psychiatry: Causes and solutions. *CNS Drugs*, 27:457–468, 2013.

[177] P Cuijpers, F Smit, E Bohlmeijer, S D Hollon, and G Andersson. Efficacy of cognitive-behavioural therapy and other psychological treatments for adult depression: meta-analytic study of publication bias. *Br J Psychiatry*, 196(3):173–8, 2010.

[178] A Yavchitz, I Boutron, A Bafeta, I Marroun, P Charles, J Mantz, and P Ravaud. Misrepresentation of randomized controlled trials in press releases and news coverage: a cohort study. *PLOS Med*, 9(9):e1001308, 2012.

[179] S N Goodman. Toward evidence-based medical statistics. 2: The Bayes factor. *Ann Intern Med*, 130(12):1005–1013, 1999.

[180] P Lockhart and B Guthrie. Trends in primary care antidepressant prescribing 1995 – 2007: a longitudinal population database analysis. *Br J Gen Pract*, 61(590):565–572, 2011.

[181] R Mojtabai and M Olfson. National trends in long-term use of antidepressant medications: results from the U.S. National Health and Nutrition Examination Survey. *J Clin Psychiatry*, 75(2):169–77, 2014.

[182] N McGauran, B Wieseler, J Kreis, Y-B Schüler, H Kölsch, and T Kaiser. Reporting bias in medical research - a narrative review. *Trials*, 11:37, 2010.

[183] J P A Ioannidis, M R Munafò, P Fusar-Poli, B A Nosek, and S P David. Publication and other reporting biases in cognitive sciences: detection, prevalence, and prevention. *Trends Cogn Sci*, 18(5):235 – 241, 2014.

[184] G B Emerson, W J Warme, F M Wolf, J D Heckman, R A Brand, and S S Leopold. Testing for the presence of positive-outcome bias in peer review: a randomized controlled trial. *JAMA Intern Med*, 170(21):1934–1939, 2010.

[185] A P Prayle, M N Hurley, and A R Smyth. Compliance with mandatory reporting of clinical trial results on ClinicalTrials.gov: cross sectional study. *BMJ*, 344(1):d7373, 2012.

[186] K Dwan, D G Altman, M Clarke, C Gamble, J P T Higgins, J A C Sterne, P R Williamson, and J J Kirkham. Evidence for the selective reporting of analyses and discrepancies in clinical trials: a systematic review of cohort studies of clinical trials. *PLOS Med*, 11(6):e1001666, 2014.

[187] B Bandelow, M Reitt, C Röver, S Michaelis, Y Görlich, and D Wedekind. Efficacy of treatments for anxiety disorders: a meta-analysis. *Int Clin Psychopharmacol*, 30(4):183–192, 2015.

[188] H Vartiainen and E Leinonen. Double-blind study of mirtazapine and placebo in hospitalized patients with major depression. *Eur Neuropsychopharmacol*, 4(2):145–150, 1994.

[189] R N Golden, C B Nemeroff, P McSorley, C D Pitts, and E M Dubé. Efficacy and tolerability of controlled-release and immediate-release paroxetine in the treatment of depression. *J Clin Psychiatry*, 63(7):577–84, 2002.

[190] T A Hammad, T P Laughren, and J Racoosin. Suicidality in pediatric patients treated with antidepressant drugs. *Arch Gen Psychiatry*, 63:332–339, 2006.

[191] M Stone, T P Laughren, M L Jones, M Levenson, P C Holland, A Hughes, T A Hammad, R Temple, and G Rochester. Risk of suicidality in clinical trials of antidepressants in adults: analysis of proprietary data submitted to US Food and Drug Administration. *BMJ*, 339:b2880, 2009.

[192] T Sharma, L S Guski, N Freund, and P C Gøtzsche. Suicidality and aggression during antidepressant treatment: systematic review and meta-analyses based on clinical study reports. *BMJ*, 352:i65, 2016.

[193] Y Molero, P Lichtenstein, J Zetterqvist, C H Gumpert, and S Fazel. Selective serotonin reuptake inhibitors and violent crime: a cohort study. *PLOS Med*, 12(9):e1001875, 2015.

[194] T J Moore, J Glenmullen, and C D Furberg. Prescription drugs associated with reports of violence towards others. *PLOS ONE*, 5(12):e15337, 2010.

[195] D Healy, A Herxheimer, and D B Menkes. Antidepressants and violence: problems at the interface of medicine and law. *PLOS Med*, 3(9):e372, 2006.

[196] P Saini, Yoon K Loke, Carrol Gamble, Douglas G Altman, Paula R Williamson, and Jamie J Kirkham. Selective reporting bias of harm outcomes within studies: findings from a cohort of systematic reviews. *BMJ*, 349:g6501, 2014.

[197] D M Hartung, D A Zarin, J-M Guise, M McDonagh, R Paynter, and M Helfand. Reporting discrepancies between the ClinicalTrials.gov results database and peer-reviewed publications. *Ann Intern Med*, 160(7):477–483, 2014.

[198] E Tang, P Ravaud, C Riveros, E Perrodeau, and A Dechartres. Comparison of serious adverse events posted at ClinicalTrials.gov and published in corresponding journal articles. *BMC Med*, 13:189, 2015.

[199] C Riveros, A Dechartres, E Perrodeau, R Haneef, I Boutron, and P Ravaud. Timing and completeness of trial results posted at ClinicalTrials.gov and published in journals. *PLOS Med*, 10(12):e1001566, 2013.

[200] J B Schroll, E Maund, and P C Gøtzsche. Challenges in coding adverse events in clinical trials: a systematic review. *PLOS ONE*, 7(7):e41174, 2012.

[201] A C Plint, D Moher, A Morrison, K Schulz, D G Altman, C Hill, and I Gaboury. Does the CONSORT checklist improve the quality of reports of randomised controlled trials? A systematic review. *Med J Aust*, 185(5):263–267, 2006.

[202] J P A Ioannidis, S J W Evans, P C Gøtzsche, R T O'Neill, D G Altman, K Schulz, and D Moher. Better reporting of harms in randomized trials: an extension of the CONSORT statement. *Ann Intern Med*, 141:781–788, 2004.

[203] B Hart, A Lundh, and L Bero. Effect of reporting bias on meta-analyses of drug trials: reanalysis of meta-analyses. *BMJ*, 344:d7202, 2012.

[204] K Rising, P Bacchetti, and L Bero. Reporting bias in drug trials submitted to the Food and Drug Administration: Review of publication and presentation. *PLOS Med*, 5(11):1561–1570, 2008.

[205] Lilly press release, https://web.archive.org/web/20080202133358/http://newsroom.lilly.com/ReleaseDetail.cfm?ReleaseID=287919, 2008.

[206] S M Miller, G J Naylor, M Murtagh, and G Winslow. A double-blind comparison of paroxetine and placebo in the treatment of depressed patients in a psychiatric outpatient clinic. *Acta Psychiatr Scand*, supp. 350:143–144, 1989.

[207] G I Spielmans and P I Parry. From evidence-based medicine to marketing-based medicine: evidence from internal industry documents. *J Bioeth Inq*, 7(1):13–29, 2010.

[208] R Smith. Medical journals are an extension of the marketing arm of pharmaceutical companies. *PLOS Med*, 2(5):e138, 2005.

[209] The PLoS Medicine Editors. Ghostwriting: the dirty little secret of medical publishing that just got bigger. *PLOS Med*, 6(9):e1000156, 2009.

[210] S Sismondo. Ghosts in the machine: publication planning in the medical sciences. *Soc Stud Sci*, 39(2):171–198, 2009.

[211] S Ebrahim, S Bance, A Athale, C Malachowski, and J P A Ioannidis. Meta-analyses with industry involvement are massively published and report no caveats for antidepressants. *J Clin Epidemiol*, 70:155–163, 2015.

[212] E C Settle, S M Stahl, S R Batey, J A Johnston, and J A Ascher. Safety profile of sustained-release bupropion in depression: Results of three clinical trials. *Clin Ther*, 21(3):454–463, 1999.

[213] A G Pedersen. Citalopram and suicidality in adult major depression and anxiety disorders. *Nord J Psychiatry*, 60(5):392–9, 2006.

[214] R K Bailey, C H Mallinckrodt, M M Wohlreich, J G Watkin, and J M Plewes. Duloxetine in the treatment of major depressive disorder: comparisons of safety and efficacy. *J Natl Med Assoc*, 98(3):437–447, 2006.

[215] P Bech, D K Kajdasz, and V Porsdal. Dose-response relationship of duloxetine in placebo-

controlled clinical trials in patients with major depressive disorder. *Psychopharmacology*, 188(3):273–280, 2006.

[216] S Brecht, D Kajdasz, S Ball, and Ml E Thase. Clinical impact of duloxetine treatment on sleep in patients with major depressive disorder. *Int Clin Psychopharmacol*, 23(6):317–24, 2008.

[217] S Brunton, F Wang, S B Edwards, A S Crucitti, M J Ossanna, D J Walker, and M J Robinson. Profile of adverse events with duloxetine treatment: A pooled analysis of placebo-controlled studies. *Drug Saf*, 33(5):393–407, 2010.

[218] J Cookson, I Gilaberte, D Desaiah, and D K Kajdasz. Treatment benefits of duloxetine in major depressive disorder as assessed by number needed to treat. *Int Clin Psychopharmacol*, 21(5):267–73, 2006.

[219] P L Delgado, S K Brannan, C H Mallinckrodt, P V Tran, R K McNamara, F Wang, J G Watkin, and M J Detke. Sexual functioning assessed in 4 double-blind placebo- and paroxetine-controlled trials of duloxetine for major depressive disorder. *J Clin Psychiatry*, 66(6):686–692, 2005.

[220] S Dodd, M Berk, K Kelin, Q Zhang, E Eriksson, W Deberdt, and J C Nelson. Application of the Gradient Boosted method in randomised clinical trials: Participant variables that contribute to depression treatment efficacy of duloxetine, SSRIs or placebo. *J Affect Disord*, 168:284–293, 2014.

[221] D L Dunner, D J Goldstein, C Mallinckrodt, Y Lu, and M J Detke. Duloxetine in treatment of anxiety symptoms associated with depression. *Depress Anxiety*, 18(2):53–61, 2003.

[222] David L Dunner, Deborah N D'Souza, Daniel K Kajdasz, Michael J Detke, and James M Russell. Is treatment-associated hypomania rare with duloxetine: Secondary analysis of controlled trials in non-bipolar depression. *J Affect Disord*, 87(1):115–119, 2005.

[223] D A Fishbain, M J Detke, J Wernicke, A S Chappell, and D K Kajdasz. The relationship between antidepressant and analgesic responses: findings from six placebo-controlled trials assessing the efficacy of duloxetine in patients with major depressive disorder. *Curr Med Res Opin*, 24(11):3105–3115, 2008.

[224] J Greist, R K McNamara, C H Mallinckrodt, J N Rayamajhi, and J Raskin. Incidence and duration of antidepressant-induced nausea: Duloxetine compared with paroxetine and fluoxetine. *Clin Ther*, 26(9):1446–1455, 2004.

[225] R Gueorguieva, C Mallinckrodt, and J H Krystal. Trajectories of depression severity in clinical trials of duloxetine: insights into antidepressant and placebo responses. *Arch Gen Psychiatry*, 68(12):1227–37, 2011.

[226] E Harada, A Schacht, T Koyama, L B Marangell, T Tsuji, and R Escobar. Efficacy comparison of duloxetine and SSRIs at doses approved in Japan. *Neuropsychiatr Dis Treat*, 11:115–123, 2015.

[227] J I Hudson, M M Wohlreich, D K Kajdasz, C H Mallinckrodt, J G Watkin, and O V Martynov. Safety and tolerability of duloxetine in the treatment of major depressive disorder: analysis of pooled data from eight placebo-controlled clinical trials. *Hum Psychopharmacol*, 20(5):327–341, 2005.

[228] S G Kornstein, M M Wohlreich, C H Mallinckrodt, J G Watkin, and D E Stewart. Duloxetine efficacy for major depressive disorder in male vs. female patients: Data from 7 randomized, double-blind, placebo-controlled trials. *J Clin Psychiatry*, 67(5):761–770, 2006.

[229] R Lewis-Fernandez, C Blanco, C H Mallinckrodt, M M Wohlreich, J G Watkin, and J M Plewes. Duloxetine in the treatment of major depressive disorder: comparisons of safety and efficacy in U.S. Hispanic and majority Caucasian patients. *J Clin Psychiatry*, 67(9):1379–1390, 2006.

[230] C H Mallinckrodt, D J Goldstein, M J Detke, Y Lu, J G Watkin, and P V Tran. Duloxetine: a new treatment for the emotional and physical symptoms of depression. *Prim Care Companion J Clin Psychiatry*, 5(1):19–28, 2003.

[231] C H Mallinckrodt, J Raskin, M M Wohlreich, J G Watkin, and M J Detke. The efficacy of duloxetine: a comprehensive summary of results from MMRM and LOCF_ANCOVA in eight clinical trials. *BMC Psychiatry*, 4:26, 2004.

[232] C H Mallinckrodt, J G Watkin, C Liu, M M Wohlreich, and J Raskin. Duloxetine in the

treatment of Major Depressive Disorder: a comparison of efficacy in patients with and without melancholic features. *BMC Psychiatry*, 5:1, 2005.

[233] C H Mallinckrodt, A Prakash, A C Andorn, J G Watkin, and M M Wohlreich. Duloxetine for the treatment of major depressive disorder: A closer look at efficacy and safety data across the approved dose range. *J Psychiatr Res*, 40:337–348, 2006.

[234] C H Mallinckrodt, A Prakash, J P Houston, R Swindle, M J Detke, and M Fava. Differential antidepressant symptom efficacy: Placebo-controlled comparisons of duloxetine and SSRIs (fluoxetine, paroxetine, escitalopram). *Neuropsychobiology*, 56:73–85, 2008.

[235] J C Nelson, M M Wohlreich, C H Mallinckrodt, M J Detke, J G Watkin, and J S Kennedy. Duloxetine for the treatment of major depressive disorder in older patients. *Am J Geriatr Psychiatry*, 13(3):227–35, 2005.

[236] J C Nelson, Y L Pritchett, O Martynov, J Y Yu, C H Mallinckrodt, and M J Detke. The safety and tolerability of duloxetine compared with paroxetine and placebo: a pooled analysis of 4 clinical trials. *Prim Care Companion J Clin Psychiatry*, 8(4):212–9, 2006.

[237] J C Nelson. Anxiety does not predict response to duloxetine in major depression: Results of a pooled analysis of individual patient data from 11 placebo-controlled trials. *Depress Anxiety*, 27(1):12–18, 2010.

[238] J C Nelson. Effects of baseline depression severity on remission rates with duloxetine and placebo in anxious and nonanxious patients with major depression. *J Clin Psychopharmacol*, 31(5):682–4, 2011.

[239] J C Nelson, Q Zhang, K Kelin, E Eriksson, W Deberdt, and M Berk. Baseline patient characteristics associated with placebo remission and their impact on remission with duloxetine and selected SSRI antidepressants. *Curr Med Res Opin*, 29(7):827–833, 2013.

[240] C B Nemeroff, A F Schatzberg, D J Goldstein, M J Detke, C Mallinckrodt, Y L Pritchett, and P V Tran. Duloxetine for the treatment of major depressive disorder. *Psychopharmacol Bull*, 36(4):106–32, 2002.

[241] D G Perahia, D K Kajdasz, D Desaiah, and P M Haddad. Symptoms following abrupt discontinuation of duloxetine treatment in patients with major depressive disorder. *J Affect Disord*, 89(1-3):207–212, 2005.

[242] D G S Perahia, D K Kajdasz, M G Royer, D J Walker, and J Raskin. Duloxetine in the treatment of major depressive disorder: an assessment of the relationship between outcomes and episode characteristics. *Int Clin Psychopharmacol*, 21(5):285–95, 2006.

[243] Y L Pritchett, M D Marciniak, P K Corey-Lisle, R A Berzon, D Desaiah, and M J Detke. Use of effect size to determine optimal dose of duloxetine in major depressive disorder. *J Psychiatr Res*, 41:311–318, 2007.

[244] A Schacht, P Gorwood, P Boyce, A Schaffer, and H Picard. Depression symptom clusters and their predictive value for treatment outcomes: results from an individual patient data meta-analysis of duloxetine trials. *J Psychiatr Res*, 53(1):54–61, 2014.

[245] R C Shelton, A C Andorn, C H Mallinckrodt, M M Wohlreich, J Raskin, J G Watkin, and Michael J Detke. Evidence for the efficacy of duloxetine in treating mild, moderate, and severe depression. *Int Clin Psychopharmacol*, 22:348–355, 2007.

[246] D E Stewart, M M Wohlreich, C H Mallinckrodt, J G Watkin, and S G Kornstein. Duloxetine in the treatment of major depressive disorder: Comparisons of safety and tolerability in male and female patients. *J Affect Disord*, 94(1-3):183–189, 2006.

[247] M E Thase, P V Tran, C Wiltse, B A Pangallo, C Mallinckrodt, and M J Detke. Cardiovascular profile of duloxetine, a dual reuptake inhibitor of serotonin and norepinephrine. *J Clin Psychopharmacol*, 25(2):132–40, 2005.

[248] M E Thase, Y L Pritchett, M J Ossanna, R W Swindle, J Xu, and M J Detke. Efficacy of duloxetine and selective serotonin reuptake inhibitors: comparisons as assessed by remission rates in patients with major depressive disorder. *J Clin Psychopharmacol*, 27(6):672–6, 2007.

[249] L Viktrup, B A Pangallo, M J Detke, and N R Zinner. Urinary side effects of duloxetine in the treatment of depression and stress urinary incontinence. *Prim Care Companion J Clin Psychiatry*, 6(2):65–73, 2004.

[250] J Wernicke, A Lledo, J Raskin, D K Kajdasz, and F Wang. An evaluation of the cardiovascular safety profile of duloxetine. *Drug Saf*, 30(5):437–455, 2007.

[251] T N Wise, D G S Perahia, B A Pangallo, W G Losin, and C G Wiltse. Effects of the

antidepressant duloxetine on body weight: analyses of 10 clinical studies. *Prim Care Companion J Clin Psychiatry*, 8(5):269–78, 2006.
[252] D S Baldwin, E H Reines, C Guiton, and E Weiller. Escitalopram therapy for major depression and anxiety disorders. *Ann Pharmacother*, 41(10):1583–1592, 2007.
[253] B Bandelow, H F Andersen, and O T Dolberg. Escitalopram in the treatment of anxiety symptoms associated with depression. *Depress Anxiety*, 24:53–61, 2007.
[254] K Demyttenaere, H F Andersen, and E H Reines. Impact of escitalopram treatment on Quality of Life Enjoyment and Satisfaction Questionnaire scores in major depressive disorder and generalized anxiety disorder. *Int Clin Psychopharmacol*, 23(5):276–286, 2008.
[255] K Demyttenaere, E H Reines, S L Lönn, and M Lader. Satisfaction with medication is correlated with outcome but not persistence in patients treated with placebo, escitalopram, or serotonin-norepinephrine reuptake inhibitors: a post hoc analysis. *The primary care companion for CNS disorders*, 13(4), 2011.
[256] J M Gorman, A Korotzer, and G Su. Efficacy comparison of escitalopram and citalopram in the treatment of major depressive disorder: pooled analysis of placebo-controlled trials. *CNS Spectrums*, 7(4 suppl 1):40–44, 2002.
[257] S H Kennedy, H F Andersen, and M E Thase. Escitalopram in the treatment of major depressive disorder: A meta-analysis. *Curr Med Res Opin*, 25(1):161–175, 2009.
[258] C D Kilts, A G Wade, H F Andersen, and T E Schlaepfer. Baseline severity of depression predicts antidepressant drug response relative to escitalopram. *Expert Opin Pharmacother*, 10(6):927–936, 2009.
[259] M Lader, H F Andersen, and T Bækdal. The effect of escitalopram on sleep problems in depressed patients. *Hum Psychopharmacol*, 20(5):349–354, 2005.
[260] R W Lam and H F Andersen. The influence of baseline severity on efficacy of escitalopram and citalopram in the treatment of major depressive disorder: An extended analysis. *Pharmacopsychiatry*, 39(5):180–184, 2006.
[261] P M Llorca, J M Azorin, N Despiegel, and P Verpillat. Efficacy of escitalopram in patients with severe depression: A pooled analysis. *Int J Clin Pract*, 59(3):268–275, 2005.
[262] G I Papakostas and K Larsen. Testing anxious depression as a predictor and moderator of symptom improvement in major depressive disorder during treatment with escitalopram. *Eur Arch Psychiatry Clin Neurosci*, 261:147–156, 2011.
[263] A G Pedersen. Escitalopram and suicidality in adult depression and anxiety. *Int Clin Psychopharmacol*, 20:139–143, 2005.
[264] Dan J Stein, David S Baldwin, Ornah T Dolberg, Nicolas Despiegel, and Borwin Bandelow. Which factors predict placebo response in anxiety disorders and major depression? An analysis of placebo-controlled studies of escitalopram. *J Clin Psychiatry*, 67(11):1741–6, nov 2006.
[265] D J Stein and A G Lopez. Effects of escitalopram on sleep problems in patients with major depression or generalized anxiety disorder. *Adv Ther*, 28(11):1021–1037, 2011.
[266] A Wade and H Friis Andersen. The onset of effect for escitalopram and its relevance for the clinical management of depression. *Curr Med Res Opin*, 22(11):2101–2110, 2006.
[267] P Bech. Meta-analysis of placebo-controlled trials with mirtazapine using the core items of the Hamilton Depression Scale as evidence of a pure antidepressive effect in the short-term treatment of major depression. *Int J Neuropsychopharmacol*, 4:337–345, 2001.
[268] J Fawcett and R L Barbin. A meta-analysis of eight randomized, double-blind, controlled clinical trials of mirtazapine for the treatment of patients with major depression and symptoms of anxiety. *J Clin Psychiatry*, 59(3):123–127, 1998.
[269] S Stahl, M Zivkov, P E Reimitz, J Panagides, and W Hoff. Meta-analysis of randomized, double-blind, placebo-controlled, efficacy and safety studies of mirtazapine versus amitriptyline in major depression. *Acta Psychiatr Scand*, 96(suppl 391):22–30, 1997.
[270] D J Carpenter, R Fong, J E Kraus, J T Davies, C Moore, and M E Thase. Meta-analysis of efficacy and treatment-emergent suicidality in adults by psychiatric indication and age subgroup following initiation of paroxetine therapy: A complete set of randomized placebo-controlled trials. *J Clin Psychiatry*, 72(11):1503–1514, 2011.
[271] G C Dunbar, J B Cohn, L F Fabre, J P Feighner, R R Fieve, J Mendels, and R K Shrivastava. A comparison of paroxetine, imipramine and placebo in depressed out-patients. *Br J Psychiatry*, 159:394–8, 1991.

[272] J P Feighner. A double-blind comparison of paroxetine, imipramine and placebo in depressed outpatients. *Int Clin Psychopharmacol*, 6 Suppl 4(suppl 4):31–35, 1992.

[273] J P Feighner, J B Cohn, L F Fabre, R R Fieve, J Mendels, R K Shrivastava, and G C Dunbar. A study comparing paroxetine placebo and imipramine in depressed patients. *J Affect Disord*, 28(2):71–9, 1993.

[274] J E Kraus, J P Horrigan, D J Carpenter, R Fong, P S Barrett, and J T Davies. Clinical features of patients with treatment-emergent suicidal behavior following initiation of paroxetine therapy. *J Affect Disord*, 120:40–47, 2010.

[275] S A Montgomery. The advantages of paroxetine in different subgroups of depression. *Int Clin Psychopharmacol*, 6(suppl 4):91–100, 1992.

[276] D L Dunner, A Lipschitz, C D Pitts, and J T Davies. Efficacy and tolerability of controlled-release paroxetine in the treatment of severe depression: post hoc analysis of pooled data from a subset of subjects in four double-blind clinical trials. *Clin Ther*, 27(12):1901–11, 2005.

[277] C Berti, D P Doogan, N R Scott, and T G Dinan. Sertraline in the treatment of depressive disorders with associated anxiety. *J Serotonin Res*, 3:151–170, 1995.

[278] C Fisch and S B Knoebel. Electrocardiographic findings in sertraline depression trials. *Drug Investigation*, 4(4):305–312, 1992.

[279] P Danjou and D Hackett. Safety and tolerance profile of venlafaxine. *Int Clin Psychopharmacol*, 10 Suppl 2:15–20, 1995.

[280] R Entsuah, G V Upton, and R Rudolph. Efficacy of venlafaxine treatment in depressed patients with psychomotor retardation or agitation: a meta-analysis. *Hum Psychopharmacol*, 10:195–200, 1995.

[281] A R Entsuah, R L Rudolph, and R Chitra. Effectiveness of venlafaxine treatment in a broad spectrum of depressed patients: a meta-analysis. *Psychopharmacol Bull*, 31(4):759–66, 1995.

[282] A R Entsuah, H Huang, and M E Thase. Response and remission rates in different subpopulations with major depressive disorder administered venlafaxine, selective serotonin reuptake inhibitors, or placebo. *J Clin Psychiatry*, 62(11):869–877, 2001.

[283] R Entsuah and B Gao. Global benefit-risk evaluation of antidepressant action: comparison of pooled data for venlafaxine, SSRIs, and placebo. *CNS spectrums*, 7(12):882–888, 2002.

[284] R D Gibbons, C H Brown, K Hur, J Davis, and J J Mann. Suicidal thoughts and behavior with antidepressant treatment: reanalysis of the randomized placebo-controlled studies of fluoxetine and venlafaxine. *Arch Gen Psychiatry*, 69(6):580–7, 2012.

[285] R Mallick, J Chen, A R Entsuah, and A F Schatzberg. Depression-free days as a summary measure of the temporal pattern of response and remission in the treatment of major depression: A comparison of venlafaxine, selective serotonin reuptake inhibitors, and placebo. *J Clin Psychiatry*, 64(3):321–330, 2003.

[286] J Mendlewicz. Pharmacologic profile and efficacy of venlafaxine. *Int Clin Psychopharmacol*, 10(suppl 2):5–13, 1995.

[287] C B Nemeroff, R Entsuah, I Benattia, M Demitrack, D M Sloan, and M E Thase. Comprehensive analysis of remission (COMPARE) with venlafaxine versus SSRIs. *Biol Psychiatry*, 63(4):424–434, 2008.

[288] R L Rudolph, R Entsuah, and R Chitra. A meta-analysis of the effects of venlafaxine on anxiety associated with depression. *J Clin Psychopharmacol*, 18(2):136–44, 1998.

[289] C Shelton, R Entsuah, S K Padmanabhan, and P E Vinall. Venlafaxine XR demonstrates higher rates of sustained remission compared to fluoxetine, paroxetine or placebo. *Int Clin Psychopharmacol*, 20(4):233–238, 2005.

[290] P H Silverstone, R Entsuah, and D Hackett. Two items on the Hamilton Depression rating scale are effective predictors of remission: comparison of selective serotonin reuptake inhibitors with the combined serotonin/norepinephrine reuptake inhibitor, venlafaxine. *Int Clin Psychopharmacol*, 17:273–280, 2002.

[291] S M Stahl, R Entsuah, and R L Rudolph. Comparative efficacy between venlafaxine and SSRIs: a pooled analysis of patients with depression. *Biol Psychiatry*, 52(12):1166–1174, 2002.

[292] M E Thase, A R Entsuah, and R L Rudolph. Remission rates during treatment with

venlafaxine or selective serotonin reuptake inhibitors. *Br J Psychiatry*, 178(3):234–241, 2001.

[293] M E Thase, R Entsuah, M Cantillon, and S G Kornstein. Relative antidepressant efficacy of venlafaxine and SSRIs: sex-age interactions. *J Womens Health*, 14(7):609–16, 2005.

[294] M H Trivedi, G J Wan, R Mallick, J Chen, R Casciano, E C Geissler, and J M Panish. Cost and effectiveness of venlafaxine extended-release and selective serotonin reuptake inhibitors in the acute phase of outpatient treatment for major depressive disorder. *J Clin Psychopharmacol*, 24(5):497–506, 2004.

[295] A R Hariri, V S Mattay, A Tessitore, B S Kolachana, F Fera, D Goldman, M F Egan, and D R Weinberger. Serotonin transporter genetic variation and the response of the human amygdala. *Science*, 297(5580):400–403, 2002.

[296] M R Munafò, S M Brown, and A R Hariri. Serotonin transporter (5-HTTLPR) genotype and amygdala activation: a meta-analysis. *Biol Psychiatry*, 63(9):852–857, 2008.

[297] S E Murphy, R Norbury, B R Godlewska, P J Cowen, Z M Mannie, C J Harmer, and M R Munafò. The effect of the serotonin transporter polymorphism (5-HTTLPR) on amygdala function: a meta-analysis. *Mol Psychiatry*, 18(4):512–20, 2013.

[298] J A Bastiaansen, M N Servaas, J B C Marsman, J Ormel, I M Nolte, H Riese, and A Aleman. Filling the gap: relationship between the serotonin-transporter-linked polymorphic region and amygdala activation. *Psychol Sci*, 25(11):2058–2066, 2014.

[299] R Bogdan, Luke W. Hyde, and A R Hariri. A neurogenetics approach to understanding individual differences in brain, behavior, and risk for psychopathology. *Mol Psychiatry*, 18(3):288–99, 2013.

[300] I-U Park, M W Peacey, and M R Munafò. Modelling the effects of subjective and objective decision making in scientific peer review. *Nature*, 506(7486):93–6, 2014.

[301] P F Sullivan, M C Neale, and K S Kendler. Genetic epidemiology of major depression: review and meta-analysis. *Am J Psychiatry*, 157:1552–1562, 2000.

[302] C Hammen. Stress and depression. *Annual Review of Clinical Psychology*, 1:293–319, 2005.

[303] A Caspi, A R Hariri, A Holmes, R Uher, and T E Moffitt. Genetic sensitivity to the environment: the case of the serotonin transporter gene and its implications for studying complex diseases and traits. *Am J Psychiatry*, 167:509–527, 2010.

[304] A Caspi, K Sugden, T E Moffitt, A Taylor, I W Craig, H L Harrington, J McClay, J Mill, J Martin, A W Braithwaite, and R Poulton. Influence of life stress on depression: moderation by a polymorphism in the 5-HTT gene. *Science*, 301(5631):386–9, 2003.

[305] C F Sharpley, S K A Palanisamy, N S Glyde, P W Dillingham, and L L Agnew. An update on the interaction between the serotonin transporter promoter variant (5-HTTLPR), stress and depression, plus an exploration of non-confirming findings. *Behav Brain Res*, 273:89–105, 2014.

[306] K Karg, M Burmeister, K Shedden, and S Sen. The serotonin transporter promoter variant (5-HTTLPR), stress, and depression meta-analysis revisited: evidence of genetic moderation. *Arch Gen Psychiatry*, 68(5):444–54, 2011.

[307] N Risch, R Herrell, T Lehner, KY Liang, L Eaves, J Hoh, A Griem, M Kovacs, J Ott, and K R Merikangas. Interaction between the serotonin transporter gene (5-HTTLPR), stressful life events, and risk of depression. *JAMA*, 301(23):2462–2471, 2009.

[308] Marcus R Munafò, Caroline Durrant, Glyn Lewis, and Jonathan Flint. Gene X environment interactions at the serotonin transporter locus. *Biol Psychiatry*, 65(3):211–9, feb 2009.

[309] L E Duncan and M C Keller. A critical review of the first 10 years of candidate gene-by-environment interaction research in psychiatry. *Am J Psychiatry*, 168:1041–1049, 2011.

[310] S Zammit, M J Owen, and G Lewis. Misconceptions about gene-environment interactions in psychiatry. *Evid Based Ment Health*, 13(3):65–68, 2010.

[311] V E Heininga, A J Oldehinkel, R Veenstra, and E Nederhof. I just ran a thousand analyses: benefits of multiple testing in understanding equivocal evidence on gene-environment interactions. *PLOS ONE*, 10(5):e0125383, 2015.

[312] P de Jonge, H J Conradi, B D Thombs, J G M Rosmalen, H Burger, and J Ormel. Prevention of false positive findings in observational studies: registration will not work but replication might. *Journal of Epidemiology and Community Health*, 65(2):95–96, 2011.

[313] A-S Jannot, T Agoritsas, A Gayet-Ageron, and T V Perneger. Citation bias favoring statistically significant studies was present in medical research. *J Clin Epidemiol*, 66:296–301, 2013.

[314] L L Kjaergard and C Gluud. Citation bias of hepato-biliary randomized clinical trials. *J Clin Epidemiol*, 55:407–410, 2002.

[315] J A Bastiaansen, Y A de Vries, and M R Munafò. Citation distortions in the literature on the serotonin-transporter-linked polymorphic region and amygdala activation. *Biol Psychiatry*, 78(8):E35–36, 2015.

[316] H A Whiteford, L Degenhardt, J Rehm, A J Baxter, A J Ferrari, H E Erskine, F J Charlson, R E Norman, A D Flaxman, N Johns, R Burstein, C J L Murray, and T Vos. Global burden of disease attributable to mental and substance use disorders: findings from the Global Burden of Disease Study 2010. *Lancet*, 382(9904):1575–1586, 2013.

[317] J Martin, J Cleak, S A G Willis-Owen, Jonathan Flint, and S Shifman. Mapping regulatory variants for the serotonin transporter gene based on allelic expression imbalance. *Mol Psychiatry*, 12(5):421–422, 2007.

[318] X Hu, G Oroszi, J Chun, T L Smith, D Goldman, and M A Schuckit. An expanded evaluation of the relationship of four alleles to the level of response to alcohol and the alcoholism risk. *Alcohol Clin Exp Res*, 29(1):8–16, 2005.

[319] J R Wendland, B J Martin, M R Kruse, K-P Lesch, and D L Murphy. Simultaneous genotyping of four functional loci of human SLC6A4, with a reappraisal of 5-HTTLPR and rs25531. *Mol Psychiatry*, 11(3):224–6, 2006.

[320] F Song, S Parekh-Bhurke, L Hooper, Y K Loke, J J Ryder, A J Sutton, C B Hing, and I Harvey. Extent of publication bias in different categories of research cohorts: a meta-analysis of empirical studies. *BMC Med Res Methodol*, 9:79, 2009.

[321] B A Nosek, J R Spies, and M Motyl. Scientific Utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspect Psychol Sci*, 7(6):615–631, 2012.

[322] R C Culverhouse, L Bowes, N Breslau, J I Nurnberger, M Burmeister, D M Fergusson, M R Munafò, N L Saccone, and L J Bierut. Protocol for a collaborative meta-analysis of 5-HTTLPR, stress, and depression. *BMC Psychiatry*, 13:304, 2013.

[323] S Every-Palmer and J Howick. How evidence-based medicine is failing due to biased trials and selective publication. *J Eval Clin Pract*, 20(6):908–914, 2014.

[324] J Barth, T Munder, H Gerger, E Nüesch, S Trelle, H Znoj, P Jüni, and P Cuijpers. Comparative efficacy of seven psychotherapeutic interventions for patients with depression: a network meta-analysis. *PLOS Med*, 10(5):e1001454, 2013.

[325] Michal Kicinski. How does under-reporting of negative and inconclusive results affect the false-positive rate in meta-analysis? A simulation study. *BMJ Open*, 4(8):e004831, 2014.

[326] M W Otto and A A Nierenberg. Assay sensitivity, failed clinical trials, and the conduct of science. *Psychother Psychosom*, 71(5):241–3, 2002.

[327] Y A de Vries, A M Roest, M Franzen, M R Munafò, and J A Bastiaansen. Citation bias and selective focus on positive findings in the literature on the serotonin transporter gene (5-HTTLPR), life stress and depression. *Psychol Med*, 46(14):2971 – 2979, 2016.

[328] J Flint, P Cuijpers, J Horder, S L Koole, and M R Munafò. Is there an excess of significant findings in published studies of psychotherapy for depression? *Psychol Med*, 45(02):439–446, 2015.

[329] D Rennie. Trial registration: a great idea switches from ignored to irresistible. *JAMA*, 292(11):1359–1362, 2004.

[330] C D DeAngelis, J M Drazen, F A Frizelle, C Haug, J Hoey, R Horton, S Kotzin, C Laine, A Marusic, A J P M Overbeke, T V Schroeder, H C Sox, and M B Van Der Weyden. Clinical trial registration: a statement from the International Committee of Medical Journal Editors. *JAMA*, 292(11):1363–1364, 2004.

[331] H Knüppel, C Metz, J J Meerpohl, and D Strech. How psychiatry journals support the unbiased translation of clinical research. A cross-sectional study of editorial policies. *PLOS ONE*, 8(10):e75995, 2013.

[332] S L Harriman and J Patel. When are clinical trials registered? An analysis of prospective versus retrospective registration. *Trials*, 17(1):187, 2016.

[333] J S Ross, G K Mulvey, E M Hines, S E Nissen, and H M Krumholz. Trial publication after

registration in ClinicalTrials.gov: A cross-sectional analysis. *PLOS Med*, 6(9):e1000144, 2009.
[334] C W Jones, L G Keil, W C Holland, M C Caughey, and T F Platts-Mills. Comparison of registered and published outcomes in randomized controlled trials: a systematic review. *BMC Med*, 13(1):282, 2015.
[335] Landelijke Stuurgroep Multidisciplinaire Richtlijnontwikkeling in de GGZ. Multidisciplinaire richtlijn: Addendum depressie bij jeugd. 2009.
[336] National Institute for Health and Care Excellence. Depression in children and young people: Identification and management in primary, community and secondary care (2015 update). 2005.
[337] Kinderformularium, https://kinderformularium.nl.
[338] Kenniscentrum KJP, http://www.kenniscentrum-kjp.nl.
[339] S E Hetrick, J E McKenzie, G R Cox, M B Simmons, and S N Merry. Newer generation antidepressants for depressive disorders in children and adolescents. *Cochrane Database Syst Rev*, (11):CD004851, 2012.
[340] C J Whittington, T Kendall, P Fonagy, D Cottrell, A Cotgrove, and E Boddington. Selective serotonin reuptake inhibitors in childhood depression: systematic review of published versus unpublished data. *Lancet*, 363(9418):1341–1345, 2004.
[341] M Miller, S A Swanson, D Azrael, V Pate, and T Stürmer. Antidepressant dose, age, and the risk of deliberate self-harm. *JAMA Intern Med*, 174(6):899, 2014.
[342] C Coupland, T Hill, R Morriss, A Arthur, M Moore, and J Hippisley-Cox. Antidepressant use and risk of suicide and attempted suicide or self harm in people aged 20 to 64: cohort study using a primary care database. *BMJ*, 350:h517, 2015.
[343] H-C Steinhausen. Recent international trends in psychotropic medication prescriptions for children and adolescents. *Eur Child Adolesc Psychiatry*, 24(6):635–640, 2015.
[344] H-C Steinhausen and C Bisgaard. Nationwide time trends in dispensed prescriptions of psychotropic medication for children and adolescents in Denmark. *Acta Psychiatr Scand*, 129(3):221–31, 2014.
[345] K R Merikangas, J-P He, J Rapoport, B Vitiello, and M Olfson. Medication use in US youth with mental disorders. *JAMA Pediatr*, 167(2):141–148, feb 2013.
[346] L P M M Wijlaars, I Nazareth, and I Petersen. Trends in depression and antidepressant prescribing in children and adolescents: a cohort study in The Health Improvement Network (THIN). *PLOS ONE*, 7(3):e33181, 2012.
[347] V Chirdkiatgumchai, H Xiao, B K Fredstrom, R E Adams, J N Epstein, S S Shah, W B Brinkman, R S Kahn, and T E Froehlich. National trends in psychotropic medication use in young children: 1994-2009. *Pediatrics*, 132(4):615–23, 2013.
[348] J F Hernandez, A K Mantel-Teeuwisse, G J M W van Thiel, S V Belitser, Jan Warmerdam, V de Valk, J A M Raaijmakers, and T Pieters. A 10-year analysis of the effects of media coverage of regulatory warnings on antidepressant use in The Netherlands and UK. *PLOS ONE*, 7(9):e45515, 2012.
[349] V Kovess, S Choppin, F Gao, M Pivette, M Husky, and E Leray. Psychotropic medication use in French children and adolescents. *J Child Adolesc Psychopharmacol*, 25(2):168–175, 2015.
[350] F Hoffmann, G Glaeske, and C J Bachmann. Trends in antidepressant prescriptions for children and adolescents in Germany from 2005 to 2012. *Pharmacoepidemiol Drug Saf*, 23(12):1268–72, 2014.
[351] G A Bushnell, T Stürmer, S A Swanson, A White, D Azrael, V Pate, and M Miller. Dosing of selective serotonin reuptake inhibitors among children and adults before and after the FDA black-box warning. *Psychiatr Serv*, 67(3):302–309, 2016.
[352] Zorginstituut Nederland. GIP Databank, https://www.gipdatabank.nl/.
[353] S T Visser, C C M Schuiling-Veninga, H J Bos, L T W de Jong-van den Berg, and M J Postma. The population-based prescription database IADB.nl: its development, usefulness in outcomes research and challenges. *Expert Rev Pharmacoecon Outcomes Res*, 13(3):285 – 292, 2013.
[354] World Health Organization Collaborating Centre. *Introduction to drug utilization research*. 2003.
[355] J R Strawn, J A Welge, A M Wehry, B Keeshin, and M A Rynn. Efficacy and tolerability

of antidepressants in pediatric anxiety disorders: a systematic review and meta-analysis. *Depress Anxiety*, 32:149–157, 2015.

[356] J C Ipser, D J Stein, S Hawkridge, and L Hoppe. Pharmacotherapy for anxiety disorders in children and adolescents. *Cochrane Database Syst Rev*, (3):CD005170, 2009.

[357] Food and Drug Administration. Revised recommendations for Celexa (citalopram hydrobromide) related to a potential risk of abnormal heart rhythms with high doses, 2011.

[358] S R Beach, W J Kostis, M Celano, J L Januzzi, J N Ruskin, P A Noseworthy, and J C Huffman. Meta-analysis of selective serotonin reuptake inhibitor-associated QTc prolongation. *J Clin Psychiatry*, 75(5):441–449, 2014.

[359] X Meng, C D Arcy, and R Tempier. Long-term trend in pediatric antidepressant use, 1983 - 2007: a population-based study. *Can J Psychiatry*, 59(2):89–97, 2014.

[360] A Pottegård, H Zoëga, J Hallas, and P Damkier. Use of SSRIs among Danish children: a nationwide study. *Eur Child Adolesc Psychiatry*, 23(12):1211–1218, 2014.

[361] T A Sheldon, N Cullum, D Dawson, A Lankshear, K Lowson, I Watt, P West, D Wright, and J Wright. What's the evidence that NICE guidance has been implemented? Results from a national evaluation using time series analysis, audit of patients' notes, and interviews. *BMJ*, 329(7473):999, 2004.

[362] A Wazana. Physicians and the pharmaceutical industry: is a gift ever just a gift? *JAMA*, 283(3):373–380, 2000.

[363] L J Cochrane, C A Olson, S Murray, M Dupuis, T Tooman, and S Hayes. Gaps between knowing and doing: understanding and assessing the barriers to optimal health care. *J Contin Educ Health Prof*, 27(2):94–102, 2007.

[364] J M Grimshaw, M Eccles, and M A Tetrod. Implementing clinical guidelines: Current evidence and future implications. *J Contin Educ Health Prof*, 24:S31–S37, 2004.

[365] R Grol and J Grimshaw. From best evidence to best practice: effective implementation of change in patients' care. *Lancet*, 362(9391):1225–30, 2003.

[366] S Michie and M Johnston. Changing clinical behaviour by making guidelines specific. *BMJ*, 328:343–345, 2004.

[367] M L M Hermens, M Oud, H Sinnema, M H Nauta, Y Stikkelbroek, D van Duin, and M Wensing. The multidisciplinary depression guideline for children and adolescents: an implementation study. *Eur Child Adolesc Psychiatry*, 24(10):1207–1218, 2015.

[368] J Grimshaw, M Eccles, R Thomas, G MacLennan, C Ramsay, C Fraser, and L Vale. Toward evidence-based quality improvement. *J Gen Intern Med*, 21(S2):S14–S20, 2006.

[369] L Forsetlund, A Bjørndal, A Rashidian, G Jamtvedt, M A O'Brien, F M Wolf, D Davis, J Odgaard-Jensen, and A D Oxman. Continuing education meetings and workshops: effects on professional practice and health care outcomes. *Cochrane Database Syst Rev*, (2):CD003030, 2009.

[370] A C Volkers, E R Heerdink, and L van Dijk. Antidepressant use and off-label prescribing in children and adolescents in Dutch general practice (2001-2005). *Pharmacoepidemiol Drug Saf*, 16:1054–1062, 2007.

[371] L I Sinclair, D M Christmas, S D Hood, J P Potokar, A Robertson, A Isaac, S Srivastava, D J Nutt, and S J C Davies. Antidepressant-induced jitteriness/anxiety syndrome: Systematic review. *Br J Psychiatry*, 194(6):483–490, 2009.

[372] GGZ Nederland. Sectorrapport GGZ 2012. Technical report, 2014.

[373] S E Bruce, K A Yonkers, M W Otto, J L Eisen, R B Weisberg, M Pagano, M T Shea, and M B Keller. Influence of psychiatric comorbidity on recovery and recurrence in generalized anxiety disorder, social phobia, and panic disorder: a 12-year prospective study. *Am J Psychiatry*, 162:1179–1187, 2005.

[374] R A Hansen, B N Gaynes, G Gartlehner, C G Moore, R Tiwari, and K N Lohr. Efficacy and tolerability of second-generation antidepressants in social anxiety disorder. *Int Clin Psychopharmacol*, 23(3):170–9, 2008.

[375] C Barbui, A Cipriani, V Patel, J L Ayuso-Mateos, and M van Ommeren. Efficacy of antidepressants and benzodiazepines in minor depression: systematic review and meta-analysis. *Br J Psychiatry*, 198(1):11–6, 2011.

[376] A Khan, A Bhat, J Faucett, R L Kolts, and W A Brown. Antidepressant-placebo differences in 16 clinical trials over 10 years at a single site: role of baseline severity. *Psychopharmacology*, 214(4):961–5, 2011.

[377] D J Stein, M B Stein, and C D Pitts. Predictors of response to pharmacotherapy in social anxiety disorder: An analysis of 3 placebo-controlled paroxetine trials. *J Clin Psychiatry*, 63(2):152–155, 2002.
[378] S A Montgomery. Implications of the severity of social phobia. *J Affect Disord*, 50(suppl 1):S17–22, 1998.
[379] M H Pollack, P Meoni, M W Otto, and D Hackett. Predictors of outcome following venlafaxine extended-release treatment of DSM-IV generalized anxiety disorder: a pooled analysis of short-and long-term studies. *J Clin Psychopharmacol*, 23(3):250–259, 2003.
[380] S B Morris and R P DeShon. Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychol Methods*, 7(1):105–125, 2002.
[381] M Adli, C Baethge, A Heinz, N Langlitz, and M Bauer. Is dose escalation of antidepressants a rational strategy after a medium-dose treatment has failed? A systematic review. *Eur Arch Psychiatry Clin Neurosci*, 255(6):387–400, 2005.
[382] J P A Ioannidis and T A Trikalinos. An exploratory test for an excess of significant findings. *Clin Trials*, 4(3):245–53, 2007.
[383] M Hoskins, J Pearce, A Bethell, L Dankova, C Barbui, W A Tol, M van Ommeren, J de Jong, S Seedat, H Chen, and J I Bisson. Pharmacotherapy for post-traumatic stress disorder: systematic review and meta-analysis. *Br J Psychiatry*, 206(2):93–100, 2015.
[384] C Allgulander, A A Dahl, C Austin, P L P Morris, J A Sogaard, R Fayyad, S P Kutcher, and C M Clary. Efficacy of sertraline in a 12-week trial for generalized anxiety disorder. *Am J Psychiatry*, 161(9):1642–1649, 2004.
[385] D S Baldwin, A K T Huusom, and E Maehlum. Escitalopram and paroxetine in the treatment of generalised anxiety disorder: randomised, placebo-controlled, double-blind study. *Br J Psychiatry*, 189(10):264–272, 2006.
[386] B Bandelow, G Chouinard, J Bobes, A Ahokas, I Eggens, S Liu, and H Eriksson. Extended-release quetiapine fumarate (quetiapine XR): a once-daily monotherapy effective in generalized anxiety disorder. Data from a randomized, double-blind, placebo- and active-controlled study. *Int J Neuropsychopharmacol*, 13(3):305–320, 2010.
[387] O Brawman-Mintzer, R G Knapp, M Rynn, R E Carter, and K Rickels. Sertraline treatment for generalized anxiety disorder: A randomized, double-blind, placebo-controlled study. *J Clin Psychiatry*, 67(6):874–881, 2006.
[388] D Hackett, V Haudiquet, and E Salinas. A method for controlling for a high placebo response rate in a comparison of venlafaxine XR and diazepam in the short-term treatment of patients with generalised anxiety disorder. *Eur Psychiatry*, 18(4):182–187, 2003.
[389] S Kasper, B Herman, G Nivoli, M Van Ameringen, A Petralia, F S Mandel, F Baldinetti, and B Bandelow. Efficacy of pregabalin and venlafaxine-XR in generalized anxiety disorder: results of a double-blind, placebo-controlled 8-week trial. *Int Clin Psychopharmacol*, 24(2):87–96, 2009.
[390] C Merideth, A J Cutler, F She, and H Eriksson. Efficacy and tolerability of extended release quetiapine fumarate monotherapy in the acute treatment of generalized anxiety disorder. *Int Clin Psychopharmacol*, 27(1):40–54, 2012.
[391] H Nicolini, D Bakish, H Duenas, M Spann, J Erickson, C Hallberg, S Ball, D Sagman, and J M Russell. Improvement of psychic and somatic symptoms in adult patients with generalized anxiety disorder: examination from a duloxetine, venlafaxine extended-release and placebo-controlled trial. *Psychol Med*, 39(2):267–276, 2009.
[392] I Nimatoudis, N P Zissis, J Kogeorgos, S Theodoropoulou, A Vidalis, and G Kaprinis. Remission rates with venlafaxine extended release in Greek outpatients with generalized anxiety disorder. A double-blind, randomized, placebo controlled study. *Int Clin Psychopharmacol*, 19(6):331–336, 2004.
[393] C Allgulander. Paroxetine in social anxiety disorder: a randomized placebo-controlled study. *Acta Psychiatr Scand*, 100(3):193–198, 1999.
[394] C Allgulander, R Mangano, J Zhang, A A Dahl, U Lepola, I Sjödin, G Emilien, H J Nyrerod, P Ostergaard, J Peltz, H Philippi, F X Poudat, S Rasmussen, G Schumann, E Tjora, and N Vaillant-Pelle. Efficacy of venlafaxine ER in patients with social anxiety disorder: A double-blind, placebo-controlled, parallel-group comparison with paroxetine. *Hum Psychopharmacol*, 19(6):387–396, 2004.
[395] S Asakura, O Tajima, and T Koyama. Fluvoxamine treatment of generalized social anxiety

disorder in Japan: a randomized double-blind, placebo-controlled study. *Int J Neuropsychopharmacol*, 10(2):263–274, 2007.

[396] D M Clark, A Ehlers, F McManus, A Hackmann, M Fennell, H Campbell, T Flower, C Davenport, and B Louis. Cognitive therapy versus fluoxetine in generalized social phobia: a randomized placebo-controlled trial. *J Consult Clin Psychol*, 71(6):1058–1067, 2003.

[397] S Kasper, D J Stein, H Loft, and R Nil. Escitalopram in the treatment of social anxiety disorder. *Br J Psychiatry*, 186(3):222–226, 2005.

[398] K A Kobak, J H Greist, J W Jefferson, and D J Katzelnick. Fluoxetine in social phobia: a double-blind, placebo-controlled pilot study. *J Clin Psychopharmacol*, 22(3):257–262, 2002.

[399] M Lader, K Stender, V Bürger, and R Nil. Efficacy and tolerability of escitalopram in 12- and 24-week treatment of social anxiety disorder: randomised, double-blind, placebo-controlled, fixed-dose study. *Depress Anxiety*, 19(4):241–8, 2004.

[400] M R Liebowitz, A J Gelenberg, and D Munjack. Venlafaxine extended release vs placebo and paroxetine in social anxiety disorder. *Arch Gen Psychiatry*, 62(2):190–198, 2005.

[401] M B Stein, A J Fyer, J R Davidson, M H Pollack, and B Wiita. Fluvoxamine treatment of social phobia (social anxiety disorder): a double-blind, placebo-controlled study. *Am J Psychiatry*, 156(5):756–760, 1999.

[402] I M van Vliet, J A den Boer, and H G Westenberg. Psychopharmacological treatment of social phobia; a double blind placebo controlled study with fluvoxamine. *Psychopharmacology*, 115(1-2):128–134, 1994.

[403] M A Jenike, S Hyman, L Baer, A Holland, W E Minichiello, L Buttolph, P Summergrad, R Seymour, and J N Ricciardi. A controlled trial of fluvoxamine in obsessive-compulsive disorder: implications for a serotonergic theory. *Am J Psychiatry*, 147(9):1209–1215, 1990.

[404] M A Jenike, L Baer, P Summergrad, W E Minichiello, A Holland, and R Seymour. Sertraline in obsessive-compulsive disorder: a double-blind comparison with placebo. *Am J Psychiatry*, 147:923–928, 1990.

[405] M A Jenike, L Baer, W E Minichiello, S L Rauch, and M L Buttolph. Placebo-controlled trial of fluoxetine and phenelzine for obsessive-compulsive disorder. *Am J Psychiatry*, 154(9):1261–1264, 1997.

[406] K Kamijima, M Murasaki, M Asai, T Higuchi, T Nakajima, C Taga, and H Matsunaga. Paroxetine in the treatment of obsessive-compulsive disorder: Randomized, double-blind, placebo-controlled study in Japanese patients. *Psychiatry Clin Neurosci*, 58:427–433, 2004.

[407] S A Montgomery, S Kasper, D J Stein, K Bang Hedegaard, and O M Lemming. Citalopram 20 mg, 40 mg and 60 mg are all effective and well tolerated compared with placebo in obsessive-compulsive disorder. *Int Clin Psychopharmacol*, 16(2):75–86, 2001.

[408] E Nakatani, A Nakagawa, T Nakao, C Yoshizato, M Nabeyama, A Kudo, K Isomura, N Kato, K Yoshioka, and M Kawamoto. A randomized controlled trial of Japanese patients with obsessive-compulsive disorder–effectiveness of behavior therapy and fluvoxamine. *Psychother Psychosom*, 74(5):269–276, 2005.

[409] J Davidson, D Baldwin, D J Stein, E Kuper, I Benattia, S Ahmed, R Pedersen, and J Musgnung. Treatment of posttraumatic stress disorder with venlafaxine extended release: a 6-month randomized controlled trial. *Arch Gen Psychiatry*, 63(10):1158–1165, 2006.

[410] R D Marshall, R Lewis-Fernandez, C Blanco, H B Simpson, S-H Lin, D Vermes, W Garcia, F Schneier, Y Neria, A Sanchez-Lacay, and M R Liebowitz. A controlled trial of paroxetine for chronic PTSD, dissociation, and interpersonal problems in mostly minority adults. *Depress Anxiety*, 24(2):77–84, 2007.

[411] F Martenyi, E B Brown, H Zhang, A Prakash, and S C Koke. Fluoxetine versus placebo in posttraumatic stress disorder. *J Clin Psychiatry*, 63:199–206, 2002.

[412] F Martenyi, E B Brown, and C D Caldwell. Failed efficacy of fluoxetine in the treatment of posttraumatic stress disorder: results of a fixed-dose, placebo-controlled study. *J Clin Psychopharmacol*, 27(2):166–170, 2007.

[413] P Tucker, R Potter-Kimball, D B Wyatt, D E Parker, C Burgin, D E Jones, and B K Masters. Can physiologic assessment and side effects tease out differences in PTSD trials?

A double-blind comparison of citalopram, sertraline, and placebo. *Psychopharmacol Bull*, 37(3):135–149, 2003.

[414] B A Van Der Kolk, J Spinazzola, M E Blaustein, J W Hopper, E K Hopper, D L Korn, and W B Simpson. A randomized clinical trial of eye movement desensitization and reprocessing (EMDR), fluoxetine, and pill placebo in the treatment of posttraumatic stress disorder: treatment effects and long-term maintenance. *J Clin Psychiatry*, 68(1):37–46, 2007.

[415] G M Asnis, F A Hameedi, A W Goddard, S G Potkin, D Black, M Jameel, K Desagani, and S W Woods. Fluvoxamine in the treatment of panic disorder: a multi-center, double-blind, placebo-controlled study in outpatients. *Psychiatry Res*, 103(1):1–14, 2001.

[416] E de Beurs, A J L M van Balkom, A Lange, P Koele, and R van Dyck. Treatment of panic disorder with agoraphobia: comparison of fluvoxamine, placebo, and psychological panic management combined with exposure and of exposure in vivo alone. *Am J Psychiatry*, 152(5):683–691, 1995.

[417] N P Nair, D Bakish, B Saxena, M Amin, G Schwartz, and T E West. Comparison of fluvoxamine, imipramine, and placebo in the treatment of outpatients with panic disorder. *Anxiety*, 2(4):192–198, 1996.

[418] M H Pollack, J J Worthington III, M W Otto, K M Maki, J W Smoller, G G Manfro, R Rudolph, and J F Rosenbaum. Venlafaxine for panic disorder: results from a double-blind, placebo-controlled study. *Psychopharmacol Bull*, 32(4):667–670, 1996.

[419] S M Stahl, I Gergel, and D Li. Escitalopram in the treatment of panic disorder: a randomized, double-blind, placebo-controlled trial. *J Clin Psychiatry*, 64:1322–1327, 2003.

[420] M B Stein, G Ron Norton, J R Walker, M J Chartier, and R Graham. Do selective serotonin re-uptake inhibitors enhance the efficacy of very brief cognitive behavioral therapy for panic disorder? A pilot study. *Psychiatry Res*, 94(3):191–200, 2000.

[421] S Leucht, S Hierl, W Kissling, M Dold, and J M Davis. Putting the efficacy of psychiatric and general medicine medication into perspective: review of meta-analyses. *Br J Psychiatry*, 200(2):97–106, 2012.

[422] H C Kraemer and D J Kupfer. Size of treatment effects and their importance to clinical research and practice. *Biol Psychiatry*, 59(11):990–6, 2006.

[423] P Cuijpers, E H Turner, S L Koole, A van Dijke, and F Smit. What is the threshold for a clinically relevant effect? The case of major depressive disorders. *Depress Anxiety*, 31(5):374–378, 2014.

[424] National Institute of Clinical Excellence. Common mental health disorders: Identification and pathways to care (clinical guideline 123). Technical report, 2011.

[425] H Baumeister. Inappropriate prescriptions of antidepressant drugs in patients with subthreshold to mild depression: time for the evidence to become practice. *J Affect Disord*, 139(3):240–3, 2012.

[426] J Spijker, R de Graaf, R V Bijl, A T F Beekman, J Ormel, and W A Nolen. Duration of major depressive episodes in the general population: results from The Netherlands Mental Health Survey and Incidence Study (NEMESIS). *Br J Psychiatry*, 181(3):208–213, 2002.

[427] S M Hendriks, J Spijker, C M M Licht, A T F Beekman, and B W J H Penninx. Two-year course of anxiety disorders: different across disorders or dimensions? *Acta Psychiatr Scand*, 128(3):212–21, 2013.

[428] National Institute of Clinical Excellence. Depression in adults: the treatment and management of depression in adults. NICE clinical guideline 90. Technical report, 2009.

[429] D Michelson, C Allgulander, K Dantendorfer, A Knezevic, D Maierhofer, V Micev, V R Paunovic, I Timotijevic, N Sarkar, L Skoglund, and S C Pemberton. Efficacy of usual antidepressant dosing regimens of fluoxetine in panic disorder: randomised, placebo-controlled trial. *Br J Psychiatry*, 179:514–518, 2001.

[430] T A Furukawa, S Z Levine, S Tanaka, Y Goldberg, M Samara, J M Davis, A Cipriani, and S Leucht. Initial severity of schizophrenia and efficacy of antipsychotics: participant-level meta-analysis of 6 placebo-controlled studies. *JAMA Psychiatry*, 72(1):14, 2015.

[431] M L Davis, J A Smits, and S G Hofmann. Update on the efficacy of pharmacotherapy for social anxiety disorder: a meta-analysis. *Expert opinion on pharmacotherapy*, 15(16):2281–2291, nov 2014.

[432] J Curtiss, L Andrews, M Davis, J Smits, and S G Hofmann. A meta-analysis of phar-

macotherapy for social anxiety disorder: an examination of efficacy, moderators, and mediators. *Expert opinion on pharmacotherapy*, 18(3):243–251, feb 2017.

[433] Y A de Vries, P de Jonge, E van den Heuvel, E H Turner, and A M Roest. Influence of baseline severity on antidepressant efficacy for anxiety disorders: Meta-analysis and meta-regression. *Br J Psychiatry*, 208(6):515–521, 2016.

[434] S G Thompson and J P T Higgins. How should meta-regression analyses be undertaken and interpreted? *Statistics in Medicine*, 21(11):1559–73, 2002.

[435] Dan J Stein, Siegfried Kasper, Elisabeth Wreford Andersen, Rico Nil, and Malcolm Lader. Escitalopram in the treatment of social anxiety disorder: Analysis of efficacy for different clinical subgroups and symptom dimensions. *Depression and Anxiety*, 20(4):175–181, 2004.

[436] Chi-Un Pae, Sheng-Min Wang, Changsu Han, Soo-Jung Lee, Ashwin A Patkar, Praksh S Masand, and Alessandro Serretti. Vortioxetine, a multimodal antidepressant for generalized anxiety disorder: a systematic review and meta-analysis. *J Psychiatr Res*, 64:88–98, may 2015.

[437] S A Montgomery, D V Sheehan, P Meoni, V Haudiquet, and D Hackett. Characterization of the longitudinal course of improvement in generalized anxiety disorder during long-term treatment with venlafaxine XR. *Journal of psychiatric research*, 36(4):209–217, 2002.

[438] D J Stein, M B Stein, W Goodwin, R Kumar, and B Hunter. The selective serotonin reuptake inhibitor paroxetine is effective in more generalized and in less generalized social anxiety disorder. *Psychopharmacology*, 158(3):267–272, nov 2001.

[439] M J Friedman, C R Marmar, D G Baker, C R Sikes, and G M Farfel. Randomized, double-blind comparison of sertraline and placebo for posttraumatic stress disorder in a Department of Veterans Affairs setting. *The Journal of clinical psychiatry*, 68(5):711–720, may 2007.

[440] Douglas G Altman and Patrick Royston. The cost of dichotomising continuous variables. *BMJ*, 332(7549):1080–1080, 2006.

[441] Clinical Study Data Request, https://clinicalstudydatarequest.com.

[442] GlaxoSmithKline. GSK Clinical Study Register, http://www.gsk-clinicalstudyregister.com.

[443] Lilly. Lilly Clinical Trial Registry, https://www.lilly.com/clinical-study-registration-and-results.

[444] GlaxoSmithKline. BRL29060A/661. Clinical evaluation of BRL29060A (paroxetine hydrochloride hydrate) in social phobia/social anxiety disorder (SAD) - a double-blind, placebo-controlled study.

[445] GlaxoSmithKline. PIR104776. Clinical evaluation of BRL29060A (paroxetine hydrochloride hydrate) in social phobia/social anxiety disorder (SAD) - a double-blind, placebo-controlled study.

[446] GlaxoSmithKline. BRL29060A/856. Clinical evaluation of BRL29060A (paroxetine hydrochloride hydrate) in generalized anxiety disorder (GAD) - a double-blind, placebo-controlled, comparative study.

[447] K T Forbush and D Watson. The structure of common and uncommon mental disorders. *Psychol Med*, 43(1):97–108, 2013.

[448] K M Keyes, N R Eaton, R F Krueger, A E Skodol, M M Wall, B Grant, L J Siever, and D S Hasin. Thought disorder in the meta-structure of psychopathology. *Psychol Med*, 43(8):1673–83, 2013.

[449] B L Rollman, B H Belnap, S Mazumdar, F Zhu, K Kroenke, H C Schulberg, and K M Shear. Symptomatic severity of PRIME-MD diagnosed episodes of panic and generalized anxiety disorder in primary care. *Journal of General Internal Medicine*, 20(7):623–8, 2005.

[450] M Katherine Shear, Timothy A Brown, David H Barlow, Roy Money, Diane E Sholomskas, Scott W Woods, Jack M Gorman, and Laszlo A Papp. Multicenter collaborative panic disorder severity scale. *Am J Psychiatry*, 154:1571–1575, 1997.

[451] H Kraemer, G A Morgan, N Leech, J A Gliner, J J Vaske, and R J Harmon. Measures of clinical significance. *The American Academy of Child and Adolescent Psychiatry*, 42(12):1524–1529, 2003.

[452] Lilly. A comparison of duloxetine hydrochloride, venlafaxine- extended release , and placebo in the treatment of generalized anxiety disorder (study HMDW).

[453] GlaxoSmithKline Clinical Study Register. A randomized, double-blind, placebo-controlled, flexible dosage trial to evaluate the efficacy and tolerability of paroxetine CR in patients with generalized anxiety disorder (GAD) (study 29060/791).
[454] GlaxoSmithKline Clinical Study Register. Paroxetine versus clomipramine and placebo in the treatment of obsessive-compulsive disorder (study 118).
[455] GlaxoSmithKline Clinical Study Register. A double-blind, multicentered, flexible-dose study of paroxetine, alprazolam and placebo in the treatment of panic disorder (study 223).
[456] M J Taylor, N Freemantle, J R Geddes, and Z Bhagwagar. Early onset of selective serotonin reuptake inhibitor antidepressant action: a systematic review and meta-analysis. *Arch Gen Psychiatry*, 63:1217 – 1223, 2006.
[457] H H Stassen, J Angst, D Hell, C Scharfetter, and A Szegedi. Is there a common resilience mechanism underlying antidepressant drug response? Evidence from 2848 patients. *J Clin Psychiatry*, 68(8):1195–1205, 2007.
[458] V Henkel, F Seemüller, M Obermeier, M Adli, M Bauer, K Kronmüller, F Holsboer, P Brieger, G Laux, W Bender, I Heuser, J Zeiler, W Gaebel, A Mayr, M Riedel, and H-J Möller. Relationship between baseline severity of depression and antidepressant treatment outcome. *Pharmacopsychiatry*, 44:27–32, 2011.
[459] J M Kim, S Y Kim, R Stewart, J A Yoo, K Y Bae, S W Jung, M S Lee, H W Yim, and T Y Jun. Improvement within 2 weeks and later treatment outcomes in patients with depressive disorders: The CRESCEND study. *J Affect Disord*, 129(1-3):183–190, 2011.
[460] H A Sackeim, S P Roose, and P W Lavori. Determining the duration of antidepressant treatment: Application of signal detection methodology and the need for duration adaptive designs (DAD). *Biol Psychiatry*, 59(6):483–492, 2006.
[461] P Gorwood, F Bayle, G Vaiva, P Courtet, E Corruble, and P-M Llorca. Is it worth assessing progress as early as week 2 to adapt antidepressive treatment strategy? Results from a study on agomelatine and a global meta-analysis. *Eur Psychiatry*, 28(6):362–371, 2013.
[462] R Uher, R H Perlis, N Henigsberg, A Zobel, M Rietschel, O Mors, J Hauser, M Z Dernovsek, D Souery, M Bajs, W Maier, K J Aitchison, A Farmer, and P McGuffin. Depression symptom dimensions as predictors of antidepressant treatment outcome: replicable evidence for interest-activity symptoms. *Psychol Med*, 42(05):967–980, 2012.
[463] J Mizushima, H Uchida, M Tada, T Suzuki, M Mimura, and S Nio. Early improvement of specific symptoms predicts subsequent recovery in bipolar depression. *J Clin Psychiatry*, 78(02):e146–e151, 2017.
[464] I Berlin and F Lavergne. Early predictors of two month response with mianserin and selective serotonin reuptake inhibitors and influence of definition of outcome on prediction. *Eur Psychiatry*, 13(3):138–142, 1998.
[465] A H Farabaugh, S Bitran, J Witte, J Alpert, S Chuzi, A J Clain, L Baer, M Fava, P J McGrath, C Dording, D Mischoulon, and G I Papakostas. Anxious depression and early changes in the HAMD-17 anxiety-somatization factor items and antidepressant treatment outcome. *Int Clin Psychopharmacol*, 25(4):214–217, 2010.
[466] S Leucht, H Fennema, R Engel, M Kaspers-Janssen, P Lepping, and A Szegedi. What does the HAMD mean? *J Affect Disord*, 148(2-3):243–248, 2013.
[467] R Tibshirani. Regression selection and shrinkage via the lasso. *J R Stat Soc Series B Stat Methodol*, 58(1):267–288, 1996.
[468] J Friedman, T Hastie, and R Tibshirani. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw*, 33(1):1–22, 2010.
[469] M A Posternak, L Baer, A A Nierenberg, and M Fava. Response rates to fluoxetine in subjects who initially show no improvement. *J Clin Psychiatry*, 72(7):949–954, 2011.
[470] T Bschor, H Kern, J Henssler, and C Baethge. Switching the antidepressant after nonresponse in adults with major depression: a systematic literature search and meta-analysis. *J Clin Psychiatry*, 77, 2016.
[471] H G Ruhé, J Huyser, J A Swinkels, and A H Schene. Dose escalation for insufficient response to standard-dose selective serotonin reuptake inhibitors in major depressive disorder - systematic review. *Br J Psychiatry*, 189:309–316, 2006.
[472] X Zhou, A V Ravindran, B Qin, C Del Giovane, Q Li, M Bauer, Y Liu, Y Fang, T da Silva,

Y Zhang, L Fang, X Wang, and P Xie. Comparative efficacy, acceptability, and tolerability of augmentation agents in treatment-resistant depression. *J Clin Psychiatry*, 76(4):e487–e498, 2015.

[473] F Hieronymus, S Nilsson, and E Eriksson. A mega-analysis of fixed-dose trials reveals dose-dependency and a rapid onset of action for the antidepressant effect of three selective serotonin reuptake inhibitors. *Translational Psychiatry*, 6(6):e834, 2016.

[474] E Jakubovski, A L Varigonda, N Freemantle, M J Taylor, and M H Bloch. Systematic review and meta-analysis: dose-response relationship of selective serotonin reuptake inhibitors in major depressive disorder. *Am J Psychiatry*, 54(7):557 – 564, 2015.

[475] S R Wisniewski, A J Rush, A A Nierenberg, B N Gaynes, D Warden, J F Luther, P J McGrath, P W Lavori, M E Thase, M Fava, and M H Trivedi. Can phase III trial results of antidepressant medications be generalized to clinical practice? A STAR*D report. *Am J Psychiatry*, 166(5):599–607, 2009.

[476] GSK Clinical Trial Registry. A phase II, placebo-controlled, double-blind study of paroxetine in depressed outpatients (29060/01/001).

[477] G C Dunbar, J L Claghorn, Ari Kiev, Karl Rickels, and W T Smith. A comparison of paroxetine and placebo in depressed outpatients. *Acta Psychiatr Scand*, 87(5):302–305, 1993.

[478] GSK Clinical Trial Registry. A double-blind comparison of paroxetine, amitriptyline, and placebo in inpatients with major depressive disorder with melancholia (29060/07/001).

[479] GSK Clinical Trial Registry. A multicenter, double-blind, placebo-controlled fixed-dose evaluation of four doses of paroxetine (29060/009).

[480] GSK Clinical Trial Registry. A multicenter, randomized, double-blind, placebo-controlled comparison of paroxetine and fluoxetine in the treatment of major depressive disorder (29060/115).

[481] GSK Clinical Trial Registry. A multicenter, randomized, double-blind, placebo-controlled comparison of paroxetine and fluoxetine in the treatment of major depressive disorder (29060/128).

[482] J G Edwards and A Goldie. Placebo-controlled trial of paroxetine in depressive illness. *Hum Psychopharmacol*, 8(3):203–209, 1993.

[483] GSK Clinical Trial Registry. A study to assess the effectiveness and tolerance of paroxetine by double-blind comparison with placebo and mianserin (29060/012-3).

[484] M H Rapaport, L S Schneider, D L Dunner, J T Davies, and C D Pitts. Efficacy of controlled-release paroxetine in the treatment of late-life depression. *J Clin Psychiatry*, 64(9):1065–1074, 2003.

[485] M H Trivedi, T A Pigotti, P Perera, K E Dillingham, M L Carfagno, and C D Pitts. Effectiveness of low doses of paroxetine controlled release in the treatment of major depressive disorder. *J Clin Psychiatry*, 65(10):1356–64, 2004.

[486] D J Goldstein, C Mallinckrodt, Y Lu, and M A Demitrack. Duloxetine in the treatment of major depressive disorder: A double-blind clinical trial. *J Clin Psychiatry*, 63(3):225–231, 2002.

[487] Lilly clinical trial registry. Duloxetine versus placebo in the treatment of major depression (HMAQ-B).

[488] Lilly clinical trial registry. Duloxetine versus placebo and paroxetine in the acute treatment of major depression (HMAT-A), 2004.

[489] D J Goldstein, Y Lu, M J Detke, C Wiltse, C Mallinckrodt, and M A Demitrack. Duloxetine in the treatment of depression: a double-blind placebo-controlled comparison with paroxetine. *J Clin Psychopharmacol*, 24(4):389–399, 2004.

[490] M J Detke, C G Wiltse, C H Mallinckrodt, R K McNamara, M A Demitrack, and I Bitter. Duloxetine in the acute and long-term treatment of major depressive disorder: A placebo- and paroxetine-controlled trial. *Eur Neuropsychopharmacol*, 14(6):457–470, 2004.

[491] D G S Perahia, F Wang, C H Mallinckrodt, D J Walker, and M J Detke. Duloxetine in the treatment of major depressive disorder: a placebo- and paroxetine-controlled trial. *Eur Psychiatry*, 21(6):367–378, sep 2006.

[492] M J Detke, Y Lu, D J Goldstein, J R Hayes, and M A Demitrack. Duloxetine, 60 mg once daily, for major depressive disorder: A randomized double-blind placebo-controlled trial. *J Clin Psychiatry*, 63(4):308–315, 2002.

[493] M J Detke, Y Lu, D J Goldstein, R K McNamara, and M A Demitrack. Duloxetine 60 mg once daily dosing versus placebo in the acute treatment of major depression. *J Psychiatr Res*, 36(6):383–390, 2002.

[494] D G S Perahia, Y L Pritchett, D K Kajdasz, M Bauer, R Jain, J M Russell, D J Walker, K A Spencer, D M Froud, J Raskin, and M E Thase. A randomized, double-blind comparison of duloxetine and venlafaxine in the treatment of patients with major depressive disorder. *J Psychiatr Res*, 42(1):22–34, 2008.

[495] A A Nierenberg, J H Greist, C H Mallinckrodt, A Prakash, A Sambunaris, G D Tollefson, and M M Wohlreich. Duloxetine versus escitalopram and placebo in the treatment of patients with major depressive disorder: onset of antidepressant action, a non-inferiority study. *Curr Med Res Opin*, 23(2):401–416, 2007.

[496] P Lee, L Shu, X Xu, C Y Wang, M S Lee, C Y Liu, J P Hong, S Ruschel, J Raskin, S A Colman, and G A Harrison. Once-daily duloxetine 60 mg in the treatment of major depressive disorder: Multicenter, double-blind, randomized, paroxetine-controlled, non-inferiority trial in China, Korea, Taiwan and Brazil. *Psychiatry Clin Neurosci*, 61(3):295–307, 2007.

[497] T M Oakes, C Katona, P Liu, M Robinson, J Raskin, and J H Greist. Safety and tolerability of duloxetine in elderly patients with major depressive disorder: a pooled analysis of two placebo-controlled studies. *Int Clin Psychopharmacol*, 28(1):1–11, 2013.

[498] T M M Oakes, A L Myers, L B Marangell, J Ahl, A Prakash, M E Thase, and S G Kornstein. Assessment of depressive symptoms and functional outcomes in patients with major depressive disorder treated with duloxetine versus placebo: Primary outcomes from two trials conducted under the same protocol. *Hum Psychopharmacol*, 27(1):47–56, 2012.

[499] GSK Clinical Trial Registry. An 8-week, randomized, double-blind, placebo-controlled, multicenter, fixed-dose study comparing the efficacy and safety of a new chemical entity (NCE) or paroxetine to placebo in moderately to severely depressed patients with major depressive disorder (NKD200006).

[500] GSK Clinical Trial Registry. A randomised, double-blind, double-dummy, parallel-group, placebo-controlled, forced dose titration study evaluating the efficacy and safety of a new chemical entity (NCE) and paroxetine in subjects with major depressive disorder (NKF100096).

[501] O A Panagiotou and J P A Ioannidis. Primary study authors of significant studies are more likely to believe that a strong association exists in a heterogeneous meta-analysis compared with methodologists. *J Clin Epidemiol*, 65(7):740–747, 2012.

[502] I Shrier, J-F Boivin, R W Platt, R J Steele, J M Brophy, F Carnevale, M J Eisenberg, A Furlan, R Kakuma, M E Macdonald, L Pilote, and M Rossignol. The interpretation of systematic reviews with meta-analyses: an objective or subjective process? *BMC Med Inform Decis Mak*, 8:19, 2008.

[503] R Dal-Ré, J S Ross, and A Marušić. Compliance with prospective trial registration guidance remained low in high-impact journals and has implications for primary end point reporting. *J Clin Epidemiol*, 75:100–107, 2016.

[504] ClinicalTrials.gov. FDAAA 801 Requirements, https://clinicaltrials.gov/ct2/manage-recs/fdaaa.

[505] M L Anderson, K Chiswell, E D Peterson, A Tasneem, J Topping, and R M Califf. Compliance with results reporting at ClinicalTrials.gov. *N Engl J Med*, 372(11):1031–9, 2015.

[506] L Hirsch. Trial registration and results disclosure: impact of US legislation on sponsors, investigators, and medical journal editors. *Curr Med Res Opin*, 24(6):1683–1689, 2008.

[507] Open Science Framework, https://osf.io/8mpji/wiki/home. Registered reports.

[508] R A Fisher. Presidential Address. *Sankhya*, 4(1):14–17, 1938.

[509] D H Marin dos Santos and A N Atallah. FDAAA legislation is working, but methodological flaws undermine the reliability of clinical trials: a cross-sectional study. *PeerJ*, 3:e1015, 2015.

[510] B Rawal and B R Deane. Clinical trial transparency: an assessment of the disclosure of results of company-sponsored trials associated with new medicines approved recently in Europe. *Curr Med Res Opin*, 30(3):395–405, 2014.

[511] Y A de Vries, E H Turner, and A M Roest. Retraction of biased journal articles. *BMJ*, 351:h5497, 2015.
[512] Restoring Study 329, https://study329.org.
[513] G M Asnis and M A Henderson. Levomilnacipran for the treatment of major depressive disorder: A review. *Neuropsychiatr Dis Treat*, 11:125–135, 2015.
[514] A Matheson. The disposable author: how pharmaceutical marketing is embraced within medicine's scholarly literature. *Hastings Cent Rep*, 46(4):31–37, 2016.
[515] J E Miller. From bad pharma to good pharma: Aligning market forces with good and trustworthy practices through accreditation, certification, and rating. *J Law Med Ethics*, 41(3):601–610, 2013.
[516] Access to Medicine Index, http://accesstomedicineindex.org.
[517] A Schafer. Biomedical conflicts of interest: a defence of the sequestration thesis-learning from the cases of Nancy Olivieri and David Healy. *J Med Ethics*, 30:8–24, 2004.
[518] F Leichsenring, A Abbass, M J Hilsenroth, F Leweke, P Luyten, J R Keefe, N Midgley, S Rabung, S Salzer, and C Steinert. Biases in research: risk factors for non-replicability in psychotherapy and pharmacotherapy research. *Psychol Med*, 47(6):1000–1011, 2016.
[519] D Van Dijk, O Manor, and L B Carey. Publication metrics and success on the academic job market. *Curr Biol*, 24(11):R516–R517, 2014.
[520] F C Fang and A Casadevall. Retracted science and the retraction index. *Infect Immun*, 79(10):3855–3859, 2011.
[521] P E Smaldino and R McElreath. The natural selection of bad science. *Royal Soc Open Sci*, 3:160384, 2016.
[522] A D Higginson and M R Munafò. Current incentives for scientists lead to underpowered studies with erroneous conclusions. *PLOS Biology*, 14(11):e2000995, 2016.
[523] S Moore, C Neylon, M P Eve, D P O'Donnell, and D Pattinson. "Excellence R Us": university research and the fetishisation of excellence. *Palgrave Communications*, 3:16105, 2017.
[524] D J Benos, E Bashari, J M Chaves, A Gaggar, N Kapoor, M LaFrance, R Mans, D Mayhew, S McGowan, A Polter, Y Qadri, S Sarfare, K Schultz, R Splittgerber, J Stephenson, Cr Tower, R G Walton, and A Zotov. The ups and downs of peer review. *Adv Physiol Educ*, 31(2):145–152, 2007.
[525] R C Culverhouse, N L Saccone, A C Horton, Y Ma, K J Anstey, and T Banaschewski et al. Collaborative meta-analysis finds no evidence of a strong interaction between stress and 5-HTTLPR genotype contributing to the development of depression. *Mol Psychiatry*, (online-first):1–10, 2017.
[526] M S Bauer. A review of quantitative studies of adherence to mental health clinical practice guidelines. *Harv Rev Psychiatry*, 10(3):138–153, 2002.
[527] R Grol, J Dalhuijsen, S Thomas, C Veld, G Rutten, and H Mokkink. Attributes of clinical guidelines that influence use of guidelines in general practice: observational study. *BMJ*, 317(7162):858–861, 1998.
[528] D Herzberg. *Happy pills in America: from Miltown to Prozac*. 2010.
[529] P D Kramer. *Listening to Prozac*. 1993.
[530] Laura A Pratt, Debra J Brody, and Q Gu. Antidepressant use in persons aged 12 and over: United States, 2005–2008. *NCHS Data Brief*, 76, 2011.
[531] P C Gotzsche. *Deadly psychiatry and organised denial*. 2015.
[532] I Kirsch. *The emperor's new drugs: exploding the antidepressant myth*. Hachette Book Group, 2010.
[533] R Whitaker. *Anatomy of an epidemic: magic bullets, psychiatric drugs, and the astonishing rise of mental illness in America*. Broadway Books, 2010.
[534] W Rief, Y Nestoriuc, A Von Lilienfeld-Toal, I Dogan, Fra Schreiber, S G Hofmann, A J Barsky, and J Avorn. Differences in adverse effect reporting in placebo groups in SSRI and tricyclic antidepressant trials: A systematic review and meta-analysis. *Drug Saf*, 32(11):1041–1056, 2009.
[535] I Kirsch. The emperor's new drugs: medication and placebo in the treatment of depression. In F Benedetti, P Enck, E Frisaldi, and M Schedlowski, editors, *Placebo*, pages 291 – 303. Springer, 2014.

[536] N C Savill, J K Buitelaar, E Anand, K A Day, T Treuer, H P Upadhyaya, and D Coghill. The efficacy of atomoxetine for the treatment of children and adolescents with attention-deficit/hyperactivity disorder: A comprehensive review of over a decade of clinical research. *CNS Drugs*, 29(2):131–151, 2015.

[537] A Aftab, C Chen, and J McBride. Flibanserin and its discontents. *Arch Womens Ment Health*, 20(2):243–247, 2017.

[538] W F Gellad, K E Flynn, and G C Alexander. Evaluation of flibanserin: science and advocacy at the FDA. *JAMA*, 314(9):869–870, 2015.

[539] R P Greenberg, R F Bornstein, M J Zborowski, S Fisher, and M D Greenberg. A meta-analysis of fluoxetine outcome in the treatment of depression. *J Nerv Ment Dis*, 182(10):547–551, 1994.

[540] M Barth, L Kriston, S Klostermann, Corrado Barbui, Andrea Cipriani, and K Linde. Efficacy of selective serotonin reuptake inhibitors and adverse events: meta-regression and mediation analysis of placebo-controlled trials. *Br J Psychiatry*, 208(2):114–119, 2016.

[541] I Bighelli, A Borghesani, and C Barbui. Is the efficacy of antidepressants in panic disorder mediated by adverse events? A mediational analysis. *PLOS ONE*, 12(6):e0178617, 2017.

[542] F Hieronymus, A Lisinski, S Nilsson, and E Eriksson. Efficacy of selective serotonin reuptake inhibitors in the absence of side effects: a mega-analysis of citalopram and paroxetine in adult depression. *Mol Psychiatry*, (online-first):1–6, 2017.

[543] P Cuijpers, A van Straten, E Bohlmeijer, S D Hollon, and G Andersson. The effects of psychotherapy for adult depression are overestimated: a meta-analysis of study quality and effect size. *Psychol Med*, 40(2):211–23, 2010.

[544] P Cuijpers, M Sijbrandij, S L Koole, G Andersson, A T F Beekman, and C F Reynolds. The efficacy of psychotherapy and pharmacotherapy in treating depressive and anxiety disorders: a meta-analysis of direct comparisons. *World Psychiatry*, 12(2):137–48, 2013.

[545] S Senn. Francis Galton and regression to the mean. *Significance*, 8(3):124–126, 2011.

[546] J Spijker, R de Graaf, R V Bijl, A T F Beekman, J Ormel, and W A Nolen. Determinants of persistence of major depressive episodes in the general population. Results from the Netherlands Mental Health Survey and Incidence Study (NEMESIS). *J Affect Disord*, 81(3):231–40, 2004.

[547] P Bower, E Kontopantelis, A J Sutton, T Kendrick, D A Richards, and S Gilbody et al. Influence of initial severity of depression on effectiveness of low intensity interventions: meta-analysis of individual patient data. *BMJ*, 346:f540, 2013.

[548] W M McDonald, I H Richard, and M R DeLong. Prevalence, etiology, and treatment of depression in Parkinson's disease. *Biol Psychiatry*, 54(3):363–375, 2003.

[549] K S Kendler, C O Gardner, and C A Prescott. Toward a comprehensive developmental model for major depression in women. *Am J Psychiatry*, 159(7):1133–1145, 2002.

[550] K S Kendler, C O Gardner, and C A Prescott. Toward a comprehensive developmental model for major depression in men. *Am J Psychiatry*, 163:115–124, 2006.

[551] A J Rush, M H Trivedi, S R Wisniewski, A A Nierenberg, J W Stewart, and D Warden et al. Acute and longer-term outcomes in depressed outpatients requiring one or several treatment steps: a STAR*D report. *Am J Psychiatry*, 163(11):1905–17, nov 2006.

[552] M H Trivedi, A J Rush, S R Wisniewski, A A Nierenberg, D Warden, L Ritz, G Norquist, R H Howland, B Lebowitz, P J McGrath, K Shores-Wilson, M M Biggs, G K Balasubramani, and M Fava. Evaluation of outcomes with citalopram for depression using measurement-based care in STAR*D: Implications for clinical practice. *Am J Psychiatry*, 163(1):28–40, 2006.

[553] G I Spielmans, M I Berman, E Linardatos, N Z Rosenlicht, A Perry, and A C Tsai. Adjunctive atypical antipsychotic treatment for major depressive disorder: a meta-analysis of depression, quality of life, and safety outcomes. *PLOS Med*, 10(3):e1001403, jan 2013.

[554] P Cuijpers, M Sijbrandij, S L Koole, G Andersson, A T F Beekman, and C F Reynolds. Adding psychotherapy to antidepressant medication in depression and anxiety disorders: a meta-analysis. *World Psychiatry*, 13(1):56–67, 2014.

[555] M A Posternak and M Zimmerman. Therapeutic effect of follow-up assessments on antidepressant and placebo response rates in antidepressant efficacy trials: meta-analysis. *Br J Psychiatry*, 190:287–92, 2007.

[556] P Skapinakis, D M Caldwell, W Hollingworth, P Bryden, N A Fineberg, P Salkovskis, N J Welton, H Baxter, D Kessler, R Churchill, and G Lewis. Pharmacological and psychotherapeutic interventions for management of obsessive-compulsive disorder in adults: a systematic review and network meta-analysis. *Lancet Psychiatry*, 3(8):730–739, 2016.

[557] A Tadic, D Wachtlin, M Berger, D F Braus, D van Calker, and N Dahmen et al. Randomized controlled study of early medication change for non-improvers to antidepressant therapy in major depression-The EMC trial. *Eur Neuropsychopharmacol*, 26(4):705–716, 2016.

[558] I Romera, V Pérez, J M Menchón, A Schacht, R Papen, D Neuhauser, M Abbar, P Svanborg, and I Gilaberte. Early switch strategy in patients with major depressive disorder. *J Clin Psychopharmacol*, 32(4):479–486, 2012.

[559] M Zimmerman, H L Clark, M D Multach, E Walsh, L K Rosenstein, and D Gazarian. Have treatment studies of depression become even less generalizable? A review of the inclusion and exclusion criteria used in placebo-controlled antidepressant efficacy trials published during the past 20 years. *Mayo Clin Proc*, 90(9):1180–1186, 2015.

[560] A N Goldstein-Piekarski, L M Williams, and K Humphreys. A trans-diagnostic review of anxiety disorder comorbidity and the impact of multiple exclusion criteria on studying clinical outcomes in anxiety disorders. *Translational Psychiatry*, 6(6):e847, 2016.

[561] Yale University Open Data Access (YODA) Project, http://yoda.yale.edu.

[562] Pfizer. Data access requests, http://www.pfizer.com/science/clinical_trials/trial_data_and_results/data_requests.

[563] Y A de Vries, A M Roest, and P de Jonge. The potential of individual patient data for research on antidepressant safety and efficacy. *Eur Neuropsychopharmacol*, 27(7):695–696, 2016.

[564] J N Jureidini and J M Nardo. Inadequacy of remote desktop interface for independent reanalysis of data from drug trials. *BMJ*, 349:g4353, 2014.

[565] National Institute of Mental Health. National Database for Clinical Trials related to Mental Illness (NDCT), https://data-archive.nimh.nih.gov/ndct.

[566] E Karyotaki, H Riper, J Twisk, A Hoogendoorn, A Kleiboer, and A Mira et al. Efficacy of self-guided internet-based cognitive behavioral therapy in the treatment of depressive symptoms: a meta-analysis of individual participant data. *JAMA Psychiatry*, 74(4):351–359, 2017.

[567] G James, D Witten, T Hastie, and R Tibshirani. *An introduction to statistical learning*. Springer, 6th edition, 2013.

[568] J Flint and K S Kendler. The genetics of major depression. *Neuron*, 81(3):484–503, 2014.

[569] G Parker. Classifying depression: Should paradigms lost be regained? *Am J Psychiatry*, 157(8):1195–1203, 2000.

Nederlandse samenvatting

Achtergrond

Angststoornissen en depressies komen veel voor. Ongeveer een kwart van de bevolking maakt gedurende het leven een depressie door, terwijl zo'n 30% een angststoornis krijgt. Ze doen zich daarnaast ook veel samen voor. Doordat deze stoornissen zoveel voorkomen en daarnaast vaak al op jonge leeftijd beginnen en een chronisch of terugkerend beloop kennen, zijn zij verantwoordelijk voor een hoge ziektelast. De behandeling van deze stoornissen komt neer op "pillen en praten": antidepressiva, psychotherapie, of de combinatie van beide. Zowel antidepressiva als psychotherapie zijn in de afgelopen decennia veelvuldig getest in gerandomiseerde, gecontroleerde studies (*randomized controlled trials*, RCTs). Ondanks deze schat aan bewijsmateriaal zijn er echter toch nog essentiële vragen onbeantwoord gebleven.

In de eerste plaats is duidelijk geworden dat de kwaliteit van de bewijslast bedreigd wordt door de aanwezigheid van *biases* (vertekeningen). Deze biases treden op wanneer bepaalde bevindingen meer kans hebben om naar buiten gebracht te worden of geciteerd te worden door andere artikelen, dan andere bevindingen. In de praktijk gaat het er dan meestal om dat positieve resultaten (die bijvoorbeeld aantonen dat een behandeling effectief of veilig is) wel gepubliceerd en geciteerd worden, terwijl negatieve bevindingen niet gepubliceerd worden, verdraaid worden zodat ze positief lijken, of ongeciteerd blijven. In 2008 is bijvoorbeeld aangetoond dat negatieve RCTs van antidepressiva voor de behandeling van depressie vaak niet waren gepubliceerd of waren gepubliceerd alsof ze positief waren. Dit heeft vermoedelijk te maken met de financiële belangen van farmaceutische bedrijven, die er baat bij hebben als hun medicijn zo effectief en veilig mogelijk lijkt zodat artsen het medicijn aan zoveel mogelijk mensen voorschrijven. Maar ook binnen de psychotherapieliteratuur is aangetoond dat studies met teleurstellende bevindingen vaker ongepubliceerd blijven. Er is echter nog weinig onderzoek gedaan naar de aanwezigheid van biases in de literatuur over angststoornissen en naar het effect van biases op de gerapporteerde veiligheid van medicatie.

In de tweede plaats is duidelijk geworden dat deze behandelingen een bescheiden effectiviteit hebben. Hoewel sommige patiënten heel goed reageren op een behandeling, moeten andere patiënten verscheidene behandelingen proberen voordat een effectieve behandeling gevonden wordt. Daarnaast lijken sommige patiënten ook heel goed te reageren op een placebo (een pil zonder werkzame bestanddelen), hetgeen suggereert dat niet iedereen een actieve behandeling nodig heeft. Bij depressieve patiënten is eerder gevonden dat vooral patiënten met relatief lichte klachten evenveel baat hebben bij een placebo als bij een antidepressivum, maar nieuwere studies spreken deze bevinding tegen en bij angststoornissen is hier tot nu toe nog nauwelijks naar gekeken. Omdat het op dit moment nog erg moeilijk is om vóór de start van de behandeling te voorspellen wie er baat zal hebben bij de behandeling, is er daarnaast ook interesse in het zo vroeg mogelijk opsporen van patiënten die niet op zullen knappen ná het starten van de behandeling. De huidige richtlijnen voor antidepressiva geven bijvoorbeeld aan dat een antidepressivum

minstens vier tot acht weken geslikt moet worden voordat het effect beoordeeld kan worden, wat een erg lange periode is voor mensen die heel depressief of angstig zijn.

Belangrijkste bevindingen

In het eerste deel van dit proefschrift beoogde ik het effect van *bias* op de bewijslast voor antidepressiva en (in mindere mate) psychotherapie in kaart te brengen en op die manier de ware effectiviteit en veiligheid op te helderen. In **hoofdstuk 2 en 3** heb ik de aanwezigheid van publicatiebias onderzocht in een cohort van studies van antidepressiva voor angststoornissen. Het bleek dat, net als bij depressie, ook bij angststoornissen de studies die vonden dat het antidepressivum effectief was een aanzienlijk grotere kans hadden om gepubliceerd te worden. Bij deze angststoornis-studies waren in werkelijkheid 72% van alle studies positief, maar 96% van de gepubliceerde studies waren positief. Wat de veiligheid betreft vond ik dat de kans op *dropout* (staken van de behandeling) wel accuraat werd weergegeven, maar dat de gepubliceerde artikelen heel weinig informatie gaven over het optreden van zeldzame maar zeer ernstige gebeurtenissen (*serious adverse events*). Wanneer wel informatie werd gegeven, klopte die informatie lang niet altijd. Zo rapporteerde een artikel dat er geen klinisch relevante gebeurtenissen (gerelateerd aan de medicatie) waren voorgekomen in de studie, terwijl twee patiënten een einde aan hun leven hadden gemaakt. Een ander artikel rapporteerde slechts een geval van "emotionele labiliteit", maar deze patiënt had tevens een suïcidepoging gedaan. Vanwege de kleine aantallen kan niet geconcludeerd worden dat deze gebeurtenissen door de medicatie veroorzaakt werden, maar dit kan ook niet uitgesloten worden. Het is belangrijk dat zulke gebeurtenissen wél gerapporteerd worden zodat een eventueel causaal verband later onderzocht kan worden door meerdere studies samen te voegen.

Het is duidelijk dat het niet publiceren van negatieve bevindingen een vertekend beeld op kan leveren. Maar ook teveel publiceren kan problematisch zijn. In **hoofdstuk 4** heb ik zogenaamde *pooled-trials publications* bestudeerd, artikelen waarin de resultaten van verschillende antidepressiva-studies gebundeld gepresenteerd worden. Dit kan soms zinvol zijn, maar het is de vraag of deze publicaties in de praktijk nuttig zijn of vooral vertekenend werken. In hoofdstuk 4 vond ik dat de onderzoeksvraag van deze publicaties vaak niet overeenkwam met de primaire onderzoeksvraag van de oorspronkelijke studies en dat de resultaten van de individuele studies zelden (3%) gepresenteerd werden. Voor de negatieve en ongepubliceerde studies waar wij naar keken, betekent dit dat de oorspronkelijke, negatieve resultaten nog steeds verstopt blijven, ook al is er over de studie gepubliceerd. Tegelijkertijd hebben bijna al deze *pooled-trials publications* positieve conclusies (bijvoorbeeld dat het medicijn effectief of veilig is of goed werkt in verschillende subgroepen van patiënten), waardoor zij bijdragen aan een literatuur die overspoeld wordt met positieve resultaten.

In **hoofdstuk 5 en 6** ging de aandacht uit naar de effecten van *spin* (positief presen-

teren van niet zo positieve resultaten) en citatiebias in een ander onderzoeksveld. In de literatuur over een variant in het serotonine-transporter-gen (5-HTTLPR) vond ik dat artikelen met negatieve bevindingen toch vaak tot positieve conclusies kwamen. Daarnaast bleek dat zowel artikelen met positieve bevindingen als positief *gepresenteerde* artikelen meer geciteerd werden dan artikelen met negatieve bevindingen én negatieve conclusies. Door deze spin en citatiebias blijven negatieve bevindingen relatief onzichtbaar, ook al zijn ze wel gepubliceerd.

In **hoofdstuk 7** bestudeerde ik hoe de effecten van *biases* zich op kunnen stapelen. Het niet-publiceren van volledige studies is inmiddels binnen de wetenschap een bekend probleem, maar hier komen nog de effecten van het niet-publiceren van bepaalde ongunstige uitkomsten (*outcome reporting bias*), spin en citatiebias bovenop. Elk van deze *biases* maakt het moeilijker om negatieve resultaten te vinden, en opgestapeld kunnen ze bijna alle negatieve resultaten onzichtbaar maken. In de literatuur over antidepressiva bij depressie vond ik bijvoorbeeld dat 52 van de 105 studies eigenlijk negatief waren, maar er waren slechts vier artikelen te vinden die helder rapporteerden dat het antidepressivum in die studie niet effectief was. Bovendien werden de positieve studies drie keer zo vaak geciteerd als negatieve studies. In de psychotherapie-literatuur deden zich vergelijkbare effecten voor.

Hoofdstuk 8 richtte zich op de vraag of de evidentie (samengevat in richtlijnen) ook echt in de praktijk wordt gebracht. Ik onderzocht hierbij of de richtlijn voor het voorschrijven van antidepressiva aan kinderen en adolescenten werd opgevolgd. Deze richtlijn schrijft voor dat de behandeling bij jongeren altijd moet starten met het specifieke antidepressivum fluoxetine, omdat voor dit middel het beste bewijs voor effectiviteit en veiligheid bij jongeren bestaat. Maar ik vond dat artsen de voorkeur gaven aan het antidepressivum citalopram, waarvan nooit is aangetoond dat het effectief is bij kinderen of adolescenten. Startdoseringen waren ook hoger dan aanbevolen, vooral bij adolescenten, die vaak een volwassen startdosering kregen. Hieruit blijkt dat de vertaalslag van de wetenschap naar de dagelijkse praktijk niet zomaar lukt, zelfs niet als er duidelijke richtlijnen bestaan.

In het tweede deel van dit proefschrift heb ik onderzocht of bepaalde klinische kenmerken, in het bijzonder ernst van de klachten en vroege verbetering in individuele symptomen, gebruikt kunnen worden om te voorspellen wie er baat zal hebben bij antidepressiva.

In **hoofdstuk 9** gebruikte ik de data van de antidepressiva-studies voor angststoornissen die in hoofdstuk 2 gebruikt waren om het effect van *bias* te onderzoeken. In dit hoofdstuk onderzocht ik of de gemiddelde effectiviteit van antidepressiva in een studie te voorspellen was aan de hand van de gemiddelde ernst van klachten in die studie. Dit bleek niet het geval te zijn. Maar het gebruik maken van gemiddelden is niet ideaal en daarom heb ik voor **hoofdstuk 10** de gegevens van individuele patiënten aangevraagd bij farmaceutische bedrijven. In dit hoofdstuk vond ik dat de ernst van klachten wél samenhing met de effectiviteit van antidepressiva voor twee angststoornissen (gegeneraliseerde angststoornis en paniekstoornis), maar niet voor drie andere angststoornissen (sociale fobie, dwangstoor-

nis, en posttraumatische stressstoornis). Dit betekent dat patiënten met een relatief lichte vorm van gegeneraliseerde angst of paniekstoornis waarschijnlijk weinig baat zullen hebben bij een antidepressivum (ten opzichte van een placebo) en andere behandelingen de voorkeur verdienen als eerste behandelstap.

In **hoofdstuk 11** heb ik onderzocht of we al na twee weken kunnen voorspellen of iemand goed zal gaan reageren op een antidepressivum. Eerdere studies hebben uitgewezen dat vroege verbetering een vrij goede voorspeller is en ik heb gekeken of we nog beter zouden kunnen voorspellen door te kijken naar vroege verbetering in specifieke symptomen en niet alleen in de totaalscore op een depressie-vragenlijst. Dit bleek maar zeer beperkt het geval te zijn, hetgeen suggereert dat artsen ongeveer evenveel informatie over de kans op een goede respons kunnen halen uit de totaalscore. Toch was dit nog niet zo'n heel goede voorspeller, omdat ook patiënten zonder vroege verbetering nog een redelijke kans hadden op een goede respons. Het zou dus veelal voorbarig zijn om al na twee weken de behandeling aan te passen bij patiënten die (nog) geen verbetering laten zien.

Conclusies

Dit proefschrift heeft laten zien hoe *bias* ons beeld van de ware effectiviteit en veiligheid van antidepressiva vertroebeld heeft. Omdat positieve resultaten vaker naar buiten gebracht worden, raakt de literatuur vertekend. Dit effect wordt nog versterkt doordat auteurs soms positievere conclusies trekken dan gerechtvaardigd door hun resultaten (*spin*) en doordat positieve resultaten ook vaker geciteerd worden. Daarnaast is in dit proefschrift een begin gemaakt met het voorspellen van een goede respons op antidepressiva, maar dit blijft voorlopig lastig. Waarschijnlijk is hiervoor behalve informatie over symptomen ook andere informatie nodig, zoals bijvoorbeeld ziektegeschiedenis of familiegeschiedenis. Doordat het gemakkelijker is geworden om individuele patiëntengegevens van gerandomiseerde studies aan te vragen voor verder onderzoek, kunnen er in de toekomst hopelijk betere voorspelmodellen ontwikkeld worden. Tezamen met een toegenomen bewustzijn van het effect van *biases* zal dit het ideaalbeeld van écht *evidence-based*, gepersonaliseerde psychiatrie dichterbij kunnen brengen.

Dankwoord

Er staat weliswaar maar één naam op de kaft van dit boekje, maar een proefschrift schrijven is toch wel echt een *team effort*. Op deze plek wil ik daarom graag de mensen bedanken die hebben bijgedragen aan mijn promotietraject.

In de eerste plaats zijn dat natuurlijk mijn promotor en copromotor. Peter en Annelieke, bedankt voor het vertrouwen en de vrijheid die jullie mij al vanaf het begin gegeven hebben. Voor jullie is het onderzoek belangrijk, maar niet belangrijker dan (het welzijn van) de onderzoekers zelf, en dat is echt niet overal in de wetenschap zo. Peter, ik moest in het begin wel een beetje wennen aan jouw poker face, maar gelukkig heb ik al snel geleerd jou wat beter te 'lezen' en jouw droge humor te waarderen. Je hebt mij in de loop der jaren vele kansen geboden, maar daarin ook mijn grenzen en wensen gerespecteerd. Annelieke, wij pasten goed bij elkaar in onze ietwat neurotische neiging om heel nauwkeurig te werken, geen overbodige eigenschap wanneer je je bezighoudt met meta-analyses. Maar een pietje precies ben je zeker niet, en onze wekelijkse overlegjes waren vooral heel plezierig. Bedankt ook voor de gezelligheid tijdens onze uitstapjes naar Wenen, Amsterdam en Boston.

Ook anderen hebben als co-auteurs een belangrijke bijdrage geleverd aan dit proefschrift, in het bijzonder Jojanneke Bastiaansen, Marcus Munafò en Erick Turner. Jojanneke, het is altijd een plezier om met jou samen te werken en ik waardeer jouw kritische maar betrokken kijk op de praktijk van het onderzoek. Marcus and Erick, thank you for great collaborations that helped to shape many of the articles in this thesis.

Ik heb het geluk gehad om tijdens mijn promotietijd een werkkamer te delen met een bonte verzameling leuke en gezellige collega's, voor de broodnodige afleiding tijdens het werk. Sanne, Nynke, Petra, Stefan, Huifang, Astrid, Maurice, Annelies en Ella, bedankt daarvoor. Dank ook aan mijn andere collega's, die het ICPE zo'n fijne plek maken om te werken.

Tot slot wil ik mijn lieve familie en vrienden bedanken. Annelene, het was enorm gezellig om de afgelopen vier jaar een werkkamer te delen (misschien niet altijd even productief) en ik zal het missen om je niet alleen als vriendin maar ook als collega en kamergenoot te hebben. Dankjewel voor je luisterend oor en je enthousiaste cheerleading. Margriet, Talita en Chris, bedankt voor de gezelligheid buiten het werk om én de discussies over ons onderzoek. Karel Jan, dankjewel voor alles: zonder jouw interesse, je zorg en je vertrouwen in mij en in mijn toekomst was dit promotietraject waarschijnlijk überhaupt nooit begonnen. Leave heit en mem, jullie snapten misschien niet altijd precies waar ik nu helemaal mee bezig was, maar waren wél altijd geïnteresseerd en trots. Leave Pier en Sonja, Harm en Laetitia, en Riejanne, we zien elkaar niet altijd even vaak, maar als we elkaar zien is het altijd goed. Het is een rijkdom om uit zo'n warm nest te komen. Riejanne, leafste suster, ook bedankt dat je mijn paranimf wou zijn, al vond je het maar een rare term. Last but not least, liefste Ate, jij hebt weliswaar alleen het staartje van dit promotietraject meegemaakt, maar dat ik jou heb ontmoet is het mooiste dat mij tijdens mijn promotie is overkomen. ♡

Curriculum vitae

Ymkje Anna de Vries (1988) was born in Westergeest, the Netherlands. After completing her secondary education (gymnasium, Lauwers College, Buitenpost) in 2005, she moved to Leiden to study Biology. During her bachelor's program, she went on exchange to the University of California, Irvine, where she subsequently worked as a research assistant before enrolling in a graduate program in Neuroscience at the University of California, San Diego. Upon her return to the Netherlands she completed a master's degree in Clinical and Psychosocial Epidemiology at the University of Groningen.

In 2014 she started her PhD program at the Interdisciplinary Center Psychopathology and Emotion regulation (ICPE, University Medical Center Groningen), under the supervision of Peter de Jonge and Annelieke Roest. She currently works as a postdoctoral researcher at the ICPE and the Developmental Psychology research group (Department of Psychology, University of Groningen).

List of publications

Roest AM, de Jonge P, Williams C, de Vries YA, Schoevers RA, Turner EH (2015). Reporting bias in clinical trials investigating the efficacy of second generation antidepressants in the treatment of anxiety disorders. JAMA Psychiatry, 72 (5), 500 - 510. DOI: 10.1001/jamapsychiatry.2015.15

Bastiaansen JA, de Vries YA, Munafò MR (2015). Citation distortions in the literature on the serotonin-transporter-linked polymorphic region and amygdala activation. Biological Psychiatry, 78 (8), E35-E36. DOI: 10.1016/j.biopsych.2014.12.007

De Vries YA, Turner EH, Roest AM (2015). Retraction of biased journal articles. BMJ, 351, h5497. DOI: 10.1136/bmj.h5497

De Vries YA, de Jonge P, van den Heuvel E, Turner EH, Roest AM (2016). Influence of baseline severity on antidepressant efficacy for anxiety disorders: meta-analysis and meta-regression. British Journal of Psychiatry, 208, 515 - 521. DOI: 10.1192/bjp.bp.115.173450

De Vries YA, de Jonge P, Kalverdijk L, Bos HJ, Schuiling-Veninga CCM, Hak E (2016). Poor guideline adherence in the initiation of antidepressant treatment in children and adolescents in the Netherlands: choice of antidepressant and dose. European Child & Adolescent Psychiatry, 25, 1161 - 1170. DOI: 10.1007/s00787-016-0836-3

Published in Dutch as: De Vries YA, de Jonge P, Kalverdijk L, Bos HJ, Schuiling-Veninga CCM, Hak E (2016). Antidepressivarichtlijnen slecht nageleefd bij jeugd. Nederlands Tijdschrift voor Geneeskunde, 160, D627.

De Vries YA, Roest AM, Franzen M, Munafò MR, Bastiaansen JA (2016). Citation bias and selective focus on positive findings in the literature on the serotonin transporter gene (5-HTTLPR), life stress and depression. Psychological Medicine, 46, 2971 - 2979. DOI: 10.1017/S0033291716000805

De Vries YA, Roest AM, Beijers L, Turner EH, de Jonge P (2016). Bias in the reporting of harms in clinical trials of second-generation antidepressants for depression and anxiety: A meta-analysis. European Neuropsychopharmacology, 26, 1752 - 1759. DOI: 10.1016/j.euroneuro.2016.09.370

De Vries YA, Roest AM, de Jonge P (2017). The potential of individual patient data for research on antidepressant safety and efficacy. European Neuropsychopharmacology, 27, 695 - 696. DOI: 10.1016/j.euroneuro.2016.11.006

De Vries YA, Roest AM, Franzen M, Munafò MR, Bastiaansen JA (2017). Moving science forward by increasing awareness of reporting and citation biases: a reply to Vrshek-Schallhorn et al. (2016). Psychological Medicine, 47, 183 - 185. DOI: 10.1017/S003329171600218X

Submitted for publication

De Vries YA, Roest AM, Turner EH, de Jonge P. Hiding negative antidepressant trials by pooling them: the pooled-trials publication bias. In revision.

De Vries YA, Roest AM, de Jonge P, Cuijpers P, Munafò MR, Bastiaansen JA. The cumulative effect of reporting and citation biases on the apparent efficacy of treatments: the case of depression. Submitted.

De Vries YA, Roest AM, Burgerhof JGM, de Jonge P. Initial severity and antidepressant efficacy for anxiety disorders: an individual patient data meta-analysis. Submitted.

De Vries YA, Roest AM, Bos EH, Burgerhof JGM, van Loo HM, de Jonge P. Predicting response to antidepressants by monitoring early improvement in individual symptoms of depression: an individual patient data meta-analysis of 8,242 patients. Submitted.